# SPATIO-TEMPORAL CRIME PREDICTION MODEL BASED ON ANALYSIS OF CRIME CLUSTERS

**A THESIS SUBMITTED TO**
**THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**
**OF**
**MIDDLE EAST TECHNICAL UNIVERSITY**

**BY**

**ESRA POLAT**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS**
**FOR**
**THE DEGREE OF MASTER OF SCIENCE**
**IN**
**GEODETIC AND GEOGRAPHIC INFORMATION TECHNOLOGIES**

**SEPTEMBER 2007**

Approval of the thesis:

**SPATIO-TEMPORAL CRIME PREDICTION MODEL BASED ON ANALYSIS OF CRIME CLUSTERS**

submitted by **ESRA POLAT** in partial fulfillment of the requirements for the degree of **Master of Science in Geodetic and Geographic Information Technologies Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Assoc. Prof. Dr. H. Şebnem Düzgün _____
Head of Department, **Geodetic and Geographic Information Technologies**

Assoc. Prof. Dr. H. Şebnem Düzgün _____
Supervisor, Geodetic and Geographic Information Technologies Dept., METU

**Examining Committee Members:**

Assoc. Prof. Dr. Oğuz Işık _____
City Planning Dept., METU

Assoc. Prof. Dr. Şebnem Düzgün _____
Mining Engineering Dept., METU

Assist. Prof. Dr. Zuhal Akyürek _____
Civil Engineering Dept., METU

Assist. Prof. Dr. Ayşegül Aksoy _____
Environmental Engineering Dept., METU

Dr. Ceylan Yozgatlıgil _____
Statistics Dept., METU

Date : 07.09.2007

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name : ESRA POLAT

Signature :

## ABSTRACT

SPATIO-TEMPORAL CRIME PREDICTION MODEL BASED ON ANALYSIS OF CRIME CLUSTERS

Polat, Esra

M.S., Department of Geodetic and Geographic Information Technologies
Supervisor: Assoc. Prof. Dr. Şebnem Düzgün

September 2007, 123 pages

Crime is a behavior disorder that is an integrated result of social, economical and environmental factors. In the world today crime analysis is gaining significance and one of the most popular subject is crime prediction. Stakeholders of crime intend to forecast the place, time, number of crimes and crime types to get precautions. With respect to these intentions, in this thesis a spatio-temporal crime prediction model is generated by using time series forecasting with simple spatial disaggregation approach in Geographical Information Systems (GIS).

The model is generated by utilizing crime data for the year 2003 in Bahçelievler and Merkez Çankaya police precincts. Methodology starts with obtaining clusters with different clustering algorithms. Then clustering methods are compared in terms of land-use and representation to select the most appropriate clustering algorithms. Later crime data is divided into daily apoch, to observe spatio-temporal distribution of crime.

In order to predict crime in time dimension a time series model (ARIMA) is fitted for each week day, Then the forecasted crime occurrences in time are disagregated according to spatial crime cluster patterns.

Hence the model proposed in this thesis can give crime prediction in both space and time to help police departments in tactical and planning operations.

Keywords: spatio-temporal crime prediction, ARIMA, simple spatial disaggregation approach, GIS, clustering.

# ÖZ

SUÇ KÜMELERİNİN ANALİZİ İLE MEKANSAL VE ZAMANSAL SUÇ TAHMİNİ MODELİNİN OLUŞTURULMASI

Polat, Esra

Y. Lisans, Jeodezi ve Coğrafi Bilgi Teknolojileri Bölümü
Tez Yöneticisi: Doç. Dr. Şebnem Düzgün

Eylül 2007, 123 sayfa

Suç sosyal, ekonomik ve çevresel faktörlerin sonuçlarının bütünleşmesi ile oluşmuş bir davranış bozukluğudur. Suç tahmini günümüzde gittikçe önemi artan suç analizi konseptinin en popüler konularından biridir. Suç ile ilgili olan paydaşlar gerekli önlemleri alabilmek için suçun yerini, zamanını, miktarını ve tipini öğrenme eğilimi içindedirler. Bu eğilimler de dikkate alınarak, bu tezde coğrafi bilgi sistemlerini (CBS) tabanlı öngörü modelleri ve basit mekansal dağıtım yaklaşımı kullanılarak zamansal ve mekansal suç tahmini modeli geliştirilmiştir.

Model 2003 yılında Bahçelievler ve Merkez Çankaya Karakol bölgelerinde meydana gelen suç verisi ile uygulanmıştır. Methodolojinin ilk adımı farklı kümeleşme algoritmaları kullanılarak suç kümeleri oluşturulmasıdır. Kümeleşme metodları arazi kullanımı ve verinin method tarafından gösterimine göre karşılaştırılmış ve en uygun kümeler seçilmiştir. Daha sonra suç verisi mekansal ve zamansal dağılımını incelemek için meydana geldiği güne göre ayrılmıştır.

Suçu zamansal olarak tahmin etmek için haftanın günleri göz önüne alınarak zaman serisi modeli (ARIMA) oluşturulmuştur. Daha sonra öngörülen suç olayları zaman ve mekansal suç kümelerine göre dağıtılmıştır.

Sonuç olarak bu tezde önerilen model zaman ve mekanda suç tahmini yapmakta ve elde edilen sonuçlar polisin taktik ve planlama operasyonlarına katkı sağlamaktadır.

Anahtar Kelimeler: zamansal ve mekansal suç tahmini, ARİMA, basit mekansal dağıtım yaklaşımı, CBS, kümeleşme.

To my family

## ACKNOWLEDGEMENTS

Firstly, I would like to thank TÜBİTAK for supporting me during my master of science period in Geodetic and Geographic Information Technologies.

I want to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Şebnem Düzgün for her guidance and confidence through the development of my thesis. Not only about the period preparation of the thesis, but also throughout our studies; she is always encouraging me to do my best. Academically, in my social life in GGIT and from now on she is really special person for me.

I want to thank all my instructors Assist. Prof. Dr. Zühal Akyürek, Assoc. Prof. Dr. Nurinnisa Usul, Assoc. Prof. Dr. Oğuz Işık, Assist Prof Dr. Ayşegül Aksoy and Dr. Ceylan Yozgatlıgil for their guidance, advice, criticism, encouragements and insight throughout the studies in GGIT and bring me to prepare this thesis.

GGIT assistants Gülcan Sarp, Arzu Erener, Pınar Aslantaş, and Serkan Kemeç for your technical assistance and great friendship that I will always miss. Thank you for everthing. Kıvanç Ertugay, Aslı Özdarıcı, Ali Özgün Ok, Dilek Koç I would also want to thank you for your assistance and friendship.

Sezen Acınan, my dear friend, you are my biggest gain in GGIT. Thank you for spending hours in project periods in laboratory, sharing your knowledge, encouracing, believing and trusting me. Also; Yalın, I want to express my deepest thanks to you being my project partner and being an exceptional for me to listen and understand me everytime.

At last, special thanks to my family Fadime and Kaya Polat supporting, encouraging me during my thesis and send their love from Bursa everytime. My brother, my home maid, my best friend Eray Polat I can not achieve without you.

**TABLE OF CONTENTS**

**LIST OF TABLES**

TABLES

## LIST OF FIGURES

FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1. Problem Definition

Crime is a behavior deviating from normal violation of the norms giving people losses and harms. Social, psychological, economical and environmental factors are to be considered in crime issue. All these concepts affect occurrence of crime in different ways. Stakeholders who have role in crime prediction are police, local governments, law enforcement agencies and people exposed to crime and offenders (Boba, 2005).

Crime is tried to be explained by various theories from different sciences. Social and psychological theories consider the root causes of crime noticing to factors such as social disorganization, personality disorders, and inadequate parenting etc. However, such theories are inefficient to be used for crime prediction, rather for explanation of it. The new advance in criminology science has important concepts that can guide crime prevention and crime analysis efforts (Boba, 2005).

Crime prevention is a significant issue that people are dealing with for centuries. Preventing crime is a necessity to make people live in more peaceful world. To achieve more calm and secure life, police is the most responsible foundation for crime prevention by targeting of resources. Police use strategic, tactical and administrative policies that assist to take precautions before an occurrence of a criminal activity. To make effective policies and improve prevention techniques, police should make use of criminological theories and crime analysis.

In the scope of this thesis clustering analyses are used to identify hot spots. Cluster analysis aims to collect data into groups according to several algorithms (Everitt, 1974). Two main groups of clustering occurring in spatial data analysis

are hierarchical and non-hierarchical/partitioning approaches. Partitioning approaches use optimization procedures to divide data into meaningful groups. K-means, fuzzy, ISODATA are the examples of partitioning approach with different algorithms. Hierarchical approaches group data set according to the type of distance specified in the algorithm. Nearest neighborhood hierarchical clustering is a type of clustering approaches. Also, there are new clustering techniques generated for specific purposes. One of them is Spatio-Temporal Analysis of Crime (STAC) which is a powerful tool to identify crime patterns and detect crime hot clusters. The other specific tool, which differing from the other algorithms by considering the underlying population when generating hot spots, is Geographic Analysis Machine (GAM).

Another issue in crime analysis is crime forecasting. Gorr and Harries (2003), indicated the purpose of crime forecast to directly support crime prevention and law enforcement. Developing highly reliable methods for forecasting future crime trends and problems is one of the most preferred ways to improve crime prevention and reduction measures. With the advance of crime forecasting, spatial and temporal predictions of crime are used to make long and short term planning. In the situation of getting accurate predictions, it is possible to manage security resources efficiently. Police give attention on highlighted areas, target patrols, allocate resources, and carry out other police interventions to prevent crime (Cohen et al., 2007).

Crime forecasting can be examined in three different concepts; spatial, temporal and spatio-temporal. Spatio-temporal crime forecasting is a new and a popular subject studied by geographers and crime analysts which includes methods from econometric modeling to neural networks (Hirsfield and Bowers, 2001).

The spatio-temporal crime prediction model used in this thesis is adapted from a study which Al-Madfai et al. (2007), generated using crime data of Lima, Ohio, USA. They used time series modeling for temporal forecasting and disaggregate predictions over generated clusters for spatial forecasting based on geographical

information systems. Resulting values were evaluated by calculating error terms and comparing disaggregation methods.

### 1.2. Objectives of the thesis

The main objective of this thesis is to develop a spatio-temporal crime prediction model based on geographical information systems coupled with spatial statistical methods. The first step in the model development is to detect spatial pattern of the crime in the study region. As crime usually occurs in the forms of clusters (Chainey and Ratcliffe, 2005), exploration of spatial pattern of crime turns out to be the detection of crime clusters. There are several algorithms for detecting crime clusters. Hence, in order to determine the most suitable clustering algorithm, K-means, Nnh hierarchical, spatio-temporal analysis of crime (STAC), fuzzy, ISODATA, and geographical analysis machine (GAM) clustering techniques are implemented. Then, crime clusters which best represents the crime pattern is chosen by relating the obtained clusters with the land use. After the detection, clusters are generated with different distance metrics; Manhattan and Euclidean. Both distance metrics are compared in terms of shape, orientation, size and area of influence. Then the next step in the crime prediction model is the prediction of number of crime in time. For this purpose, the time series model of ARIMA is used for daily crime data. Finally the predicted numbers of crime for each weekday are disaggregated to clusters with two different distance metrics. Hence by this way the developed model enables predicting spatio-temporal occurrence of crime. Finally, the predictive performance of the model is evaluated by calculating spatio-temporal root mean square error of the model.

### 1.3. Organization of the thesis

To outline the thesis, environmental crime theories and concepts in crime forecasting are explained in the second chapter. Also, crime analysis with geographical information systems is in the scope of the second chapter. The

Crime theories in environmental criminology include rational choice approach, crime pattern theory, routine activities theory, and situational crime prevention.

In the third chapter, data and study area are described, the methodology of the study and the information about methods and techniques are given. Chapter 3 explains the structure of the thesis by methodology and assisted to the reader to get knowledge about the data, area and methods.

The fourth chapter involves the analysis and comparison of clustering algorithms. The implemented clustering methods are mapped and evaluated in terms of land-use, algorithm and suitability for the crime prediction model.

The fifth chapter is the chapter where the effect of using different distance metrics in selected clustering algorithms are analyzed. The results are evaluated by assessing differences and similarities of Euclidean and Manhattan distance metrics. This chapter gives background information about the influence of distance metrics on clustering algorithms.

The concept of sixth chapter is to generate a crime prediction model with statistical model fitting and simple spatial disaggregation approach. Box-Jenkins ARIMA model is used for forecasting future crime occurrence in time and predictions are assigned to the area by simple spatial disaggregation approach. To disaggregate the data, clusters for each day are generated according to selected algorithms. Then, crime prediction model is generated and applied to the study area to indicate the results.

The last chapter is the discussion and conclusion part evaluating the results of this thesis. Clustering methods, comparison of distance metrics, efficiency of spatio-temporal crime prediction model, and contribution of this study are discussed and future studies based on this work are recommended.

# CHAPTER 2

# GENERAL VIEW OF CRIME PREDICTION AND PREVENTION MODELS

Crime is an integrated result of social, political, economical and environmental conditions that happen in a specific geography in a specific time. Why and where crime takes place is quite important to analyze the three main reasons of crime: A likely offender, a suitable target and an absence of guardian. Crime can be prevented or reduced by making people less likely to be offend, making targets less vulnerable, and by making guardians more available (Web3).

Street crimes, organized crimes, drug crimes, political and white collar crimes are main categories of crime events (Boba, 2005). All these appearing types are subdivided into different kinds of crime. For example, robbery, assault, burglary and auto theft are types of crime categorize under street crime. All categories of crime require different prevention techniques.

Stakeholders of crime mapping are police departments, politicians, non-governmental organizations, local governments, law enforcement agencies and community (NIJ, 1995). While planning crime prevention measures, stakeholders make use of crime theories to understand the spatial phenomenon in crime. Also, stakeholders benefit from geographical information systems and crime prediction models for crime prevention. In the following sections related crime theories, crime analysis with geographical information systems, and concepts in crime forecasting are explained.

## 2.1. Environmental Criminology and Crime Theories

While, traditional criminology deals with root causes of crime and why people commit an offence; the main areas of environmental criminology are crime

patterns, opportunities for committing crime, prevention of victims in criminal activities and environment prone to crime (Brantingham and Brantingham, 1881).

Felson and Clarke (1998), stated that if an offender is willing to commit a crime; he/she should pass over some physical obstacles. The probability of occurrence of crime increases with the absence of physical barriers preventing crime. An experiment was made among students comparing their behavior towards an event that was prepared before. They gave opportunity to students to be safe about cheating, lying and stealing. When the results of the study was investigated, very few of the students resisted to the given opportunities (Felson and Clarke, 1998).

The main purpose of the environmental criminology is not to detail with why specific crime incidents are committed by specific people, but to understand the behavior under supporting crime opportunities and the distribution of the environmental factors. Hence, after a criminal activity, crime analysts should search the phenomena by these questions:

- Are similar criminal activities occurring at the same place?
- Is the problem arising due to the individual's behavior?
- Is the place is prone to crime? (Boba, 2005)

The results of the questions are lightning how the environmental factors provide the opportunities for crime. There are three crime theories in environmental criminology that have interest in easy and appealing opportunities which are driving people to crime (Boba, 2005):

- Rational choice theory,
- Crime pattern theory,
- Routine activities theory.

These theories explain the behavior of an offender and a victim, the reasons of intersection of the activities, the creation of crime opportunities. Rational choice

theory handles the individual's behaviors whereas the other theories deal with the social structures. Environmental criminology explains the reasons of opportunities, the behavior of offenders and victims, and the environment under the concept of these theories.

The main idea of rational choice theory is how the offenders evaluate the choice of committing crime with respect to chances and values gained after the event. Any people might intend to commit a crime if a suitable environment and opportunities exist (Felson and Clarke, 1998).

According to rational choice theory, people do not prefer to be in a criminal activity where the risk of being catching is high and earnings are lower than expected. In crime prevention, rational choice theory is a good guide in some conditions to understand the reasons of criminal activity. Crime analysts and police utilize the theory increasing the risk for offenders to assist the crime prevention (Boba, 2005).

Rational choice theory focuses on decisions of offenders. When a criminal activity occurs, of course the aim of an offender is to gain a value after the event. According to the theory, decision of an offender becomes clearer when risks and outcomes of crime is proportioned to the earnings. However, most of the times an offender can not be able to evaluate the price objectively considering the event. Because of this reason, the rational choice theory is treated as an approach and considered with the other theories together (Chainey and Rattcliffe, 2005).

To understand choices for committing a specific crime clearly, reasons should be categorized and analyzed separately even if they belong to the same type of crime. For example, an offender can steal a car for several reasons: Auto-theft, to benefit for another criminal activity, journey, etc. All the reasons should be carefully evaluated and analyzed to be one step ahead of an offender. Rational choice theory tries to figure out the choices of offenders by looking the life from offenders view. According to the theory offenders make decisions by checking out

the losses and gaining before getting involved in an activity. In order to repeat again, offenders prefer the conditions that clearer and quicker than evaluating the steps of the activity (Boba, 2005).

According to crime pattern theory, crime generally occurs where the daily activities of both offenders and victims intersect (Felson and Clarke, 1998). Those activity areas of a victim are the places used for daily activities; whereas the same definition means for offenders known and common places. For instance, if offenders reside in crowded, demanding shopping centers; people are more probable to be exposed to a crime. Local crime patterns express if the relationship between people and physical environment create an opportunity for crime. There are three main parts of the theory:

- Nodes: Nodes are the dense criminal areas that are intersection points of routine activities.
- Roads: The dense crime areas are connected with roads which people pass to during daily activities.
- Edges: Boundaries of daily activity areas (Felson and Clarke, 1998).

The basis of environmental criminology is crime triangle (Figure 1). Crime triangle states that a criminal activity occurs when a vulnerable target and a motivated offender meet in a convenient environment (Felson and Clarke, 1998).

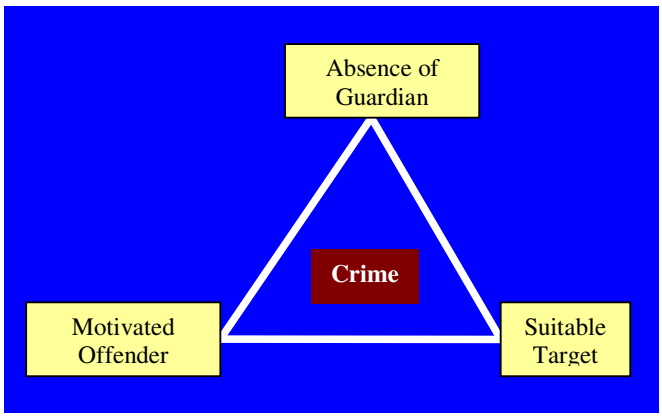



Figure 2.1. Crime Triangle

Routine activities theory focuses on how opportunities for crime change based on changes in behavior on societal level. The spreading usage of internet is one of the most recent examples for this suggestion. Internet brought several opportunities for people who are intended to commit crime. New types of crime arise while the technology, social behavior changes and so it is going to be harder to cope with every day. People responsible for security should be aware of these changes in social life and get right prevention measures (Boba, 2005).

In addition to these theories, situational crime prevention approach (SCPA) should be considered. Environmental criminology realized some of the conditions giving opportunity to crime and decided to get over the situations to prevent crime. With respect to crime theories and crime triangle, prevention measures are classified to five groups in SCPA:

- Increasing the offender's perceived effort: For example increasing the police forces, putting alarm systems, etc.
- Increasing the offender's perceived risk: For example increasing the number of street lamps, getting a mobile phone, etc.
- Reducing the offender's rewards: For example, hiding valuable thing in cars and houses, taking less money, etc.
- Reduce the offender's motivation: For example, limiting the number of people entering the facilities, preventing gangs in schools, etc.
- Removing excuses for crime: For example, putting warning boards, etc (Boba, 2005).

### 2.2. Analysis of crime with geographic information systems

Complete, consistent, and reliable sets of crime data, up-to-date information, skilled staff, appropriate geographical information systems (GIS) background and related statistical software are requirements to utilize the high technology advances in crime. Data is always one of the most important parts of the analysis. Sufficient, clear and utilizable data is the result of severely, carefully collected

and manipulated data. Geographical information system (GIS) is a computer based technology that should be applied by a professional staff to obtain satisfactory results (Hirschfield and Bowers, 2001).

Crime mapping has long been a subset of the process today known as crime analysis. Before the development of computerized crime mapping, incidents were represented by traditional maps with pins stuck in it. Since the traditional pin maps have serious limitations, crime maps are now supported with computer technology (Web3). Continual development in computer technology innovated geographical information systems (GIS) for the studies where the geography should be concerned. Many industries and organizations are the users of GIS. Crime maps started to be created with GIS to archive, manipulate and query the crime data; to update crime patterns; to make spatial analysis and to develop crime prediction and prevention models.

Crime mapping plays an important role in pro-active policing and crime prevention in stages of data collection, data evaluation and data analysis. The application areas of crime mapping are recording and mapping crime activities, predicting crime, identifying crime hot spots and patterns, monitoring the impact of crime reduction measures and communicating with stakeholders (Chainey and Ratcliffe, 2005).
Police departments are one of the most important stakeholders since they are using crime maps to take precautions against crime, to make crime analysis, to optimize resource allocation and to act the incident location as soon as possible.

Although a wide variety of statistical and analytic techniques exist to examine crime problems, analysts are increasingly using geographical information systems (GIS) and mapping software to identify areas of crime concentration (Web3). Crime maps are used to control and prevent crime in most of the developed countries. Specific crime softwares have been created to be used in crime analysis generally by different police departments. ATAC, CrimeStat, Crimeview, Geobalance, RCAGIS are some of the softwares developed for spatial analyzing

of crime in an advance manner (Boba, 2005). Spatial crime analysis softwares are capable of data entry, data manipulation, pattern identification, clustering, data mining, and geographic profiling. Analyzing crime with related softwares is complex but an advanced and reliable method to reach satisfactory results.

The first step of crime analysis with the help of crime maps is to analyze the current status of incidents. Incidents often form spatial patterns by populating in certain locations and at certain times (Chainey and Rattcliffe, 2005). Therefore, there are some relationships between the geography and crime incidents. Also, it is believed that various types of incidents occur in different geographies like street crimes in roads, and burglary in residential areas. This means that the location has different effects for each crime type.

Spatial patterns are identified to detect hot spots that are explained by Chainey and Ratcliffe (2005), as a geographical area of higher than average crime. It is an area where crime incidents densely populated compared to the average. Being aware of where hot spot occurs is important to understand the underlying reason of a criminal activity. Not only the place of a hot spot, but also the size, scale and the relationship with the other hot spots should be considered. Size and scale of a hot spot represent the identity of crime and also the reason behind the event.
Researchers and crime analysts are concerned with identifying a crime hot spot reliably and objectively. Spatial statistical techniques with geographic information systems are combined to detect real crime hot spots. Hot spot analysis using crime mapping is classified into five general techniques: (Vann and Garson, 2003)

- Visual interpretation: This is the simplest way of identifying hot spots. Analysts are creating crime maps and interpret the densely populated crime areas as a hot spot. However, this method has subjectivity that is not admitted by most of the researchers.

- Choropleth mapping: Choropleth map is simply a thematic map that uses colors to specify intensities. Choropleth mapping uses boundaries such as

cities, police boundaries to identify hot spots. One most important subject is to define the boundaries according to the usage purposes.

- Grid-cell analysis: Grid-cell analysis changes the meaning of a boundary. Uniform-sized grid cells are created and point falling within a grid is taken into account to identify hot spots. Quadrat and Kernel analysis are two examples of grid-cell analysis.

- Point pattern or cluster analysis: These analyses include statistical algorithms to define clusters of point data. Nearest neighbourhood index, k-means clustering, fuzzy clustering are some examples of point pattern and cluster analysis.

- Spatial Autocorrelation: Statistical algorithm designed to establish spatial relationships among clusters of points. Positive autocorrelation indicates clustering, whereas zero autocorrelation means random distribution of the samples.

### 2.3. Crime Forecasting

Nowadays crime forecasting is used to predict repeated actions amongst offenders, types and rates of future crimes. Using demographic and economic factors, complex statistical modeling and crime mapping or combination of these are the main methodologies that can be met in the literature for crime forecasting (Web4).

Forecasting techniques are divided into two categories in terms of predicted time period. Crime forecasting includes long-term forecast models for planning and policy applications in broader manner and short-term forecast models for tactical decision making (Gorr and Harries, 2003). Strategic planning over long-term horizon is vital in organizations in private sector for plant location and layout, demand forecasting, and planning new products. Such an issue is not compatible with the security sector. Comparing the return of long-term and short-term

forecasting in terms of crime prevention, long-term forecasts are of little value to police. Short-term forecasts are preferred for being one step ahead of offenders. Generally, one week or one month forecasts are utilized by police departments to anticipate and prevent crime (Gorr and Olligschlaeger, 1997).

Performing convenient forecasting methods according to the data is the most essential part to achieve accurate and reliable predictions. Although determining the appropriate model is significant, it should be remembered that whatever method is be applied, all these techniques are only estimation procedures and are highly dependent on what information goes into the analysis (Web4). Having insufficient data excites poorly generated forecasting models and as a result non-ideal predictions.

Actually, method used in crime forecasting is related with the questions; when, where and how much crime activity occur. According to the answer of the questions, model turns out to be spatial, temporal and spatio-temporal. In literature there are no distinct borders between these models.

Where crime will happen next is the issue of spatial forecasting methods. Several techniques are built for predicting crime patterns. Hot spot mapping, univariate techniques, multivariate techniques, and point process modeling are included in these techniques. Hot spot mapping is the simplest way of identifying future crime patterns (Chainey and Rattcliffe, 2005). Univariate and multivariate are short-term forecasting techniques. Univariate is an explorative forecast model which uses a previous value of one variable to predict crime. Multivariate is a leading indicator forecast model which uses multiple variables that affect crime to predict crime patterns. The main difference between the two types of model is that extrapolative models can only continue or extrapolate existing crime patterns into the future; whereas, leading indicator models can be able to forecast new crime patterns not yet observed. Predictions from either type of the models are utilizable resources to prevent crime (Gorr and Olligschlaeger, 1997). Chainey and Rattcliffe (2005) refer to Groff and La Vigne (2002) stating that the third technique of point process

modeling that is deals with offenders' behavior combining multivariate models and kriging.

Temporal forecasting is related with temporal distribution of crime (Chainey and Rattcliffe, 2005). There is not any model specified for temporal forecasting in literature. Temporal distribution and time series graphs of crime are the tools for predicting the probable time of event.

Gorr and Olligschlaeger (1997) stated that difficulties in obtaining reliable and accurate data and sufficient models result in few improvements in spatio-temporal crime forecasting. Several techniques are built for predicting where and when crime will happen in future. Econometric models, Box Jenkins Space-time autoregressive integrated moving average (STARIMA) multivariate transfer models, and artificial neural networks are used in spatio-temporal crime forecasting to predict crime (Gorr and Olligschlaeger, 1997).

Econometric models develop forecasts of a time series using one or more related time series and possibly past values of the time series. This approach involves developing a regression model in which the time series is forecast as the dependent variable; the related time series as well as the past values of the time series are the independent or predictor variables (Web5).

STARIMA modeling is a method for modeling univariate time series. The power of ARMA related model is that they can incorporate both autoregressive terms and moving average terms. The use of ARMA models was popularized by Box and Jenkins. Although both AR and MA models were previously known and used, Box and Jenkins provided a systematic approach for modeling both AR and MA terms in the model. ARMA models are also commonly known as Box-Jenkins models or ARIMA models (Web6).

Another spatio-temporal forecasting technique is artificial neural networks. These techniques are based on past space and time information of crime data. The neural network learns the influence of space and time patterns to crime iteratively to predict where and when crime is going to happen (Gorr and Olligschlaeger, 1997).

Deadman (2003) compared econometric and time series modeling (ARIMA) approaches and developed a number of forecasts of residential burglary in England and Wales. Gorr and Olligschlaeger (1997) introduced chaotic cellular forecasting derived from artificial neural networks as a new spatio-temporal forecasting technique.

Harries (2003) and Deadman (2003) studied long-term forecasting that made 3-year forecast of residential burglaries for England and Wales, using an econometric model as a forecasting technique. Felson and Paulson (2003), Corcoran et al. (2003) and Gorr et al. (2003) provided methods for short-term, tactical decision making forecasting; performing different categories.

# CHAPTER 3

# DESCRIPTION OF STUDY AREA, DATA AND METHODOLOGY USED IN CRIME PREDICTION MODEL

In this chapter, study area in terms of geographical, social and economical aspects, the data are defined and the methodology of adapted spatio-temporal crime prediction model is explained.

## 3.1. Description of the Study Area

The study area is located in Çankaya district of Ankara, which is the capital city of Turkey. The population of Ankara is 4 007 860 for year 2000 according to the data from Türkish Statistical Institute. Ankara has 30 715 $km^2$ area including Altındağ, Çankaya, Etimesgut, Keçiören, Mamak, Sincan, Yenimahalle, Akyurt, Ayaş, Bala, Beypazarı, Çamlıdere, Çubuk, Elmadağ, Evren, Gölbaşı, Güdül, Haymana, Kalecik, Kazan, Kızılcahamam, Nallıhan, Polatlı ve Şereflikoçhisar districts (Web1). Among these districts Çankaya includes most of the neighborhoods near to the city center.

Çankaya district is divided into 10 police precincts namely, Bahçelievler, Dikmen, Merkez, Cebeci, On Nisan, Esat, Kavaklıdere, Yıldızevler, Şehit Mustafa Düzgün and Ellici Yıl. The study area consists of two of these precincts; Bahçelievler and Merkez police station zones covering 8 $km^2$in Çankaya district. Merkez Çankaya police station zone includes 15 neighborhoods, Kızılay, Meşrutiyet, Yücetepe, Anıttepe, Eti, Maltepe, Korkut Reis, Kavaklıdere, Sağlık, Fidanlık, Namık Kemal, Cumhuriyet, Kültür, Kocatepe and Devlet; where Bahçelievler police station zone includes, Bahçelievler, yukarı Bahçelievler, Beşevler and Mebusevleri (Akpınar, 2005). Figure 3.1 illustrates the study area.

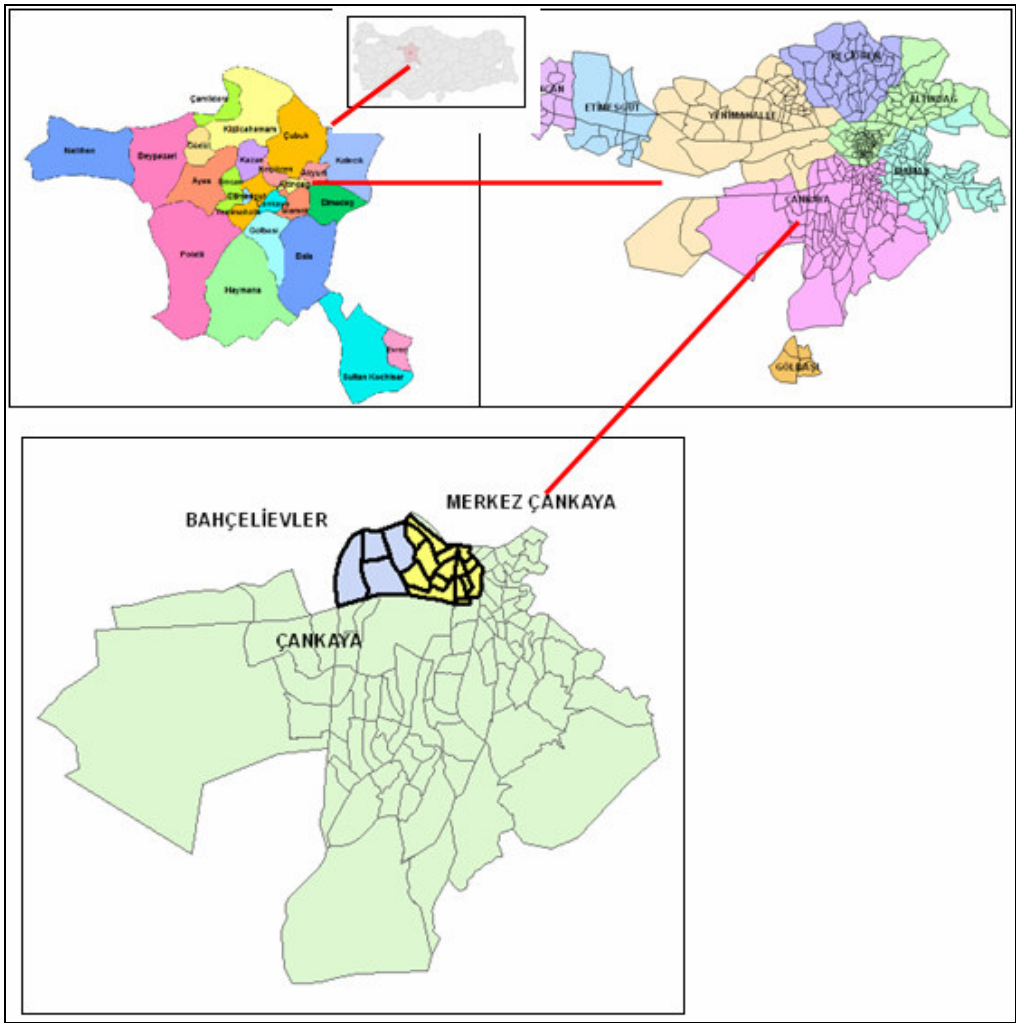


Figure 3.1.The Study Area

Having both cultural and commercial facilities located in Çankaya, it is the most developed district in Ankara. Evidently, nearly 60 percent of public associations in Turkey are in this district. Çankaya has various types of land-use areas from residential, commercial, public buildings, parks, museums to military zones, etc. It is more probable to observe different types of crime in a mixed type of land-use. For example, burglary from house is mostly seen at residential areas, where pickpocketing occurs dominantly in the commercial areas (Web2).

As stated before, Çankaya is an important place for cultural and sport activities. There are lots of cinemas, theatres, swimming pools and parks located in the

district. In fact, Atatürk museum, Anıt Park, and Municipality Ice Skating Facilities are all located in Çankaya. Also, business centers, shopping malls are densely populatedin the district which contributes the economical development of the city. Therefore, Çankaya is one of the most vital and active part of Ankara attracting offenders to commit crime (Akpınar, 2005).

Total area of Çankaya is 203 $km^2$ which is mostly residential areas. The percentage of residential areas in Çankaya is 75, where only 4% of area is commercial. Although the proportion of commercial area is significantly low, the area is very important and large relative to the other districts. The mixed usage area is 16% with both residential and commercial areas. The rest includes the education, military, public associations and cultural facilities (Akpınar, 2005).

Bahçelievler and Merkez Çankaya police station zones cover several types of land-use area. Merkez Çankaya part includes mostly commercial areas such as Kızılay Square, the heart of Ankara. Also, public associations, governmental organizations and little residential area are in the boundary of Merkez Çankaya police precinct. Bahçelievler police precinct is responsible of mostly residential areas but also important commercial areas exist in the area especially at both sides of the main streets in Bahçelievler. For example, Aşgabat Street which is known as Seventh Street is located in Bahçelievler. Anıtkabir and other museums and parks are also in Bahçelievler which makes it attractive for offenders. These two precincts are active and crowded areas, providing many opportunities for offenders.

### 3.2. Description of the Data

Crime data of year 2003 in the study area is used in the analysis which is illustrated in Figure 3.2. Data is taken from Akpınar (2005). Spatial and temporal information regarding to these incidents were obtained from Ankara Police Directorate. Crime data were recorded by two police stations Bahçelievler and Merkez Çankaya. Data include number, address, occurrence time, location and

type. Five types of data are available which are murder, usurp, burglary, auto related crimes and pickpocketing. However, in this study all the crime types are aggregated to have higher number of incidents for constructing reliable prediction model.

There are 1885 incidents recorded in Bahçelievler and Merkez Çankaya police precincts. Crime incidents are mapped with graduated symbols (Figure 3.2) which different sizes of features represent particular values of variables (Boba, 2005). To understand the density of criminal activities in the area the best way is to apply graduated symbols as incidents are overlapping.

Obtaining accurate and reliable crime data that reflect the real incidents is hardly achievable. Staffs in police departments and people exposed to crime are responsible for giving and saving the data. As stated in previous chapter, one of the necessities of computer based crime mapping is qualitative staff. When staff is not expertise in computers and content of the work, errors may exist even in entering data. Data should be properly collected and stored to get more reliable results because analyzing crime incidents is difficult as crime is a mobile phenomenon. In addition to staff, the people are not always informing police when exposed to crime. According to the interview with Rana Sampson (2006), people do not want to struggle with the bureaucracy not only in our country but in the entire world. She confirmed that the recorded crime is really a smaller part of all the criminal activities. As in Turkey, most popular crime events recorded by police are auto related crimes in USA in order to get the payment from the insurance companies.

Another data obtained from Akpınar (2005) are land-use and land marks depicted in Figure 3.3 and 3.4 including the land-use areas, major and minor roads and, land marks.

Commercial areas are larger in Merkez Çankaya police precinct, including hotels, high schools, primary schools, and car parks (Figure 3.4). Bahçelievler has

residential areas with high schools, primary schools, post offices, universities and retail centers. Land-marks were built according to the main structure of the area like hotels are mainly at Merkez Çankaya, whereas schools are located at Bahçelievler.

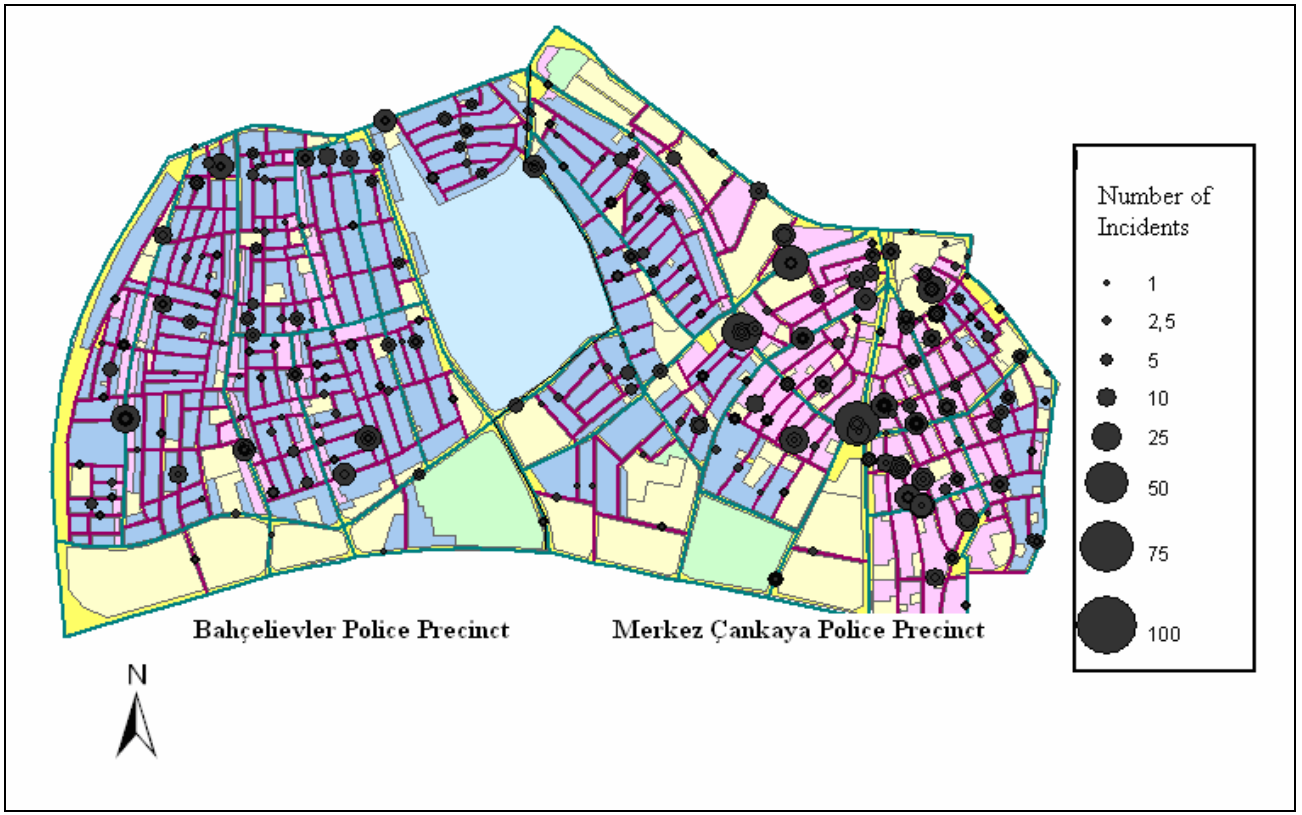



Figure 3.2.Crime incidents in study area.

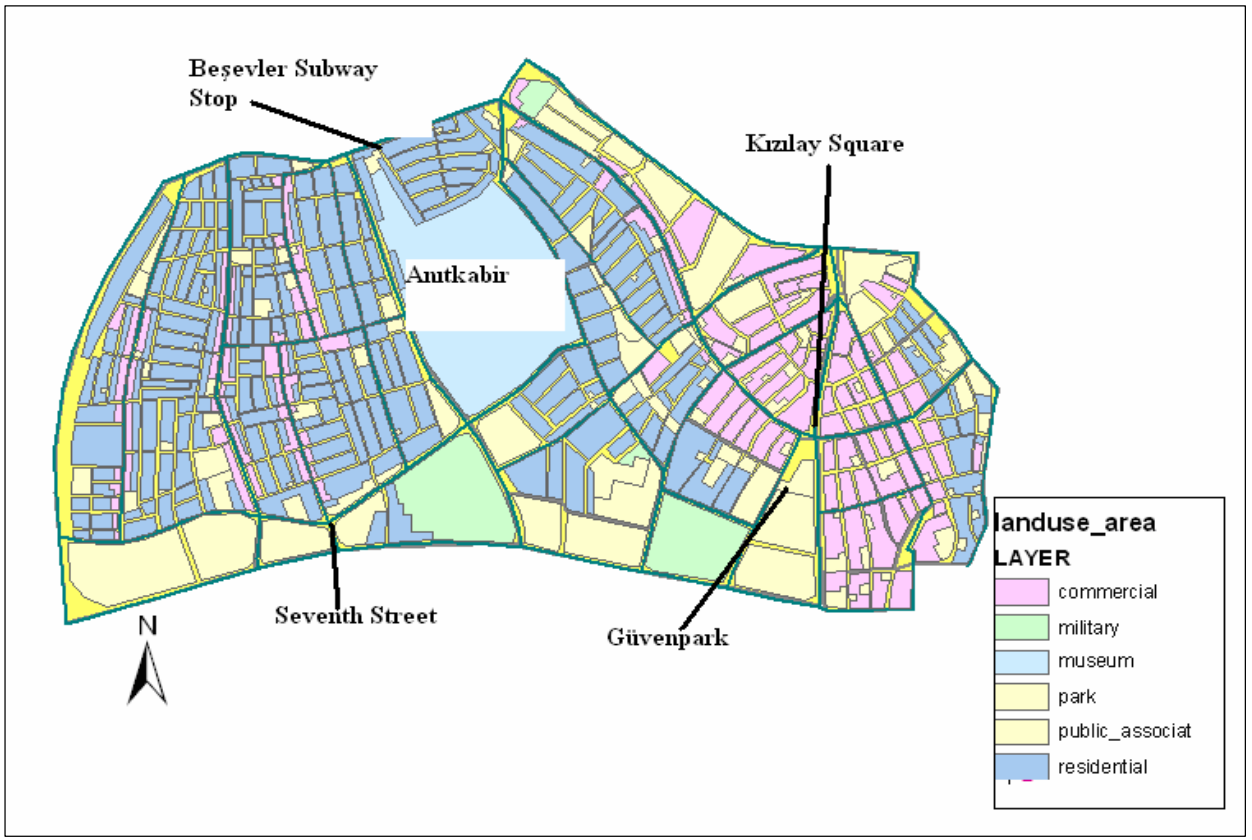



Figure 3.3.Land use of the study area.

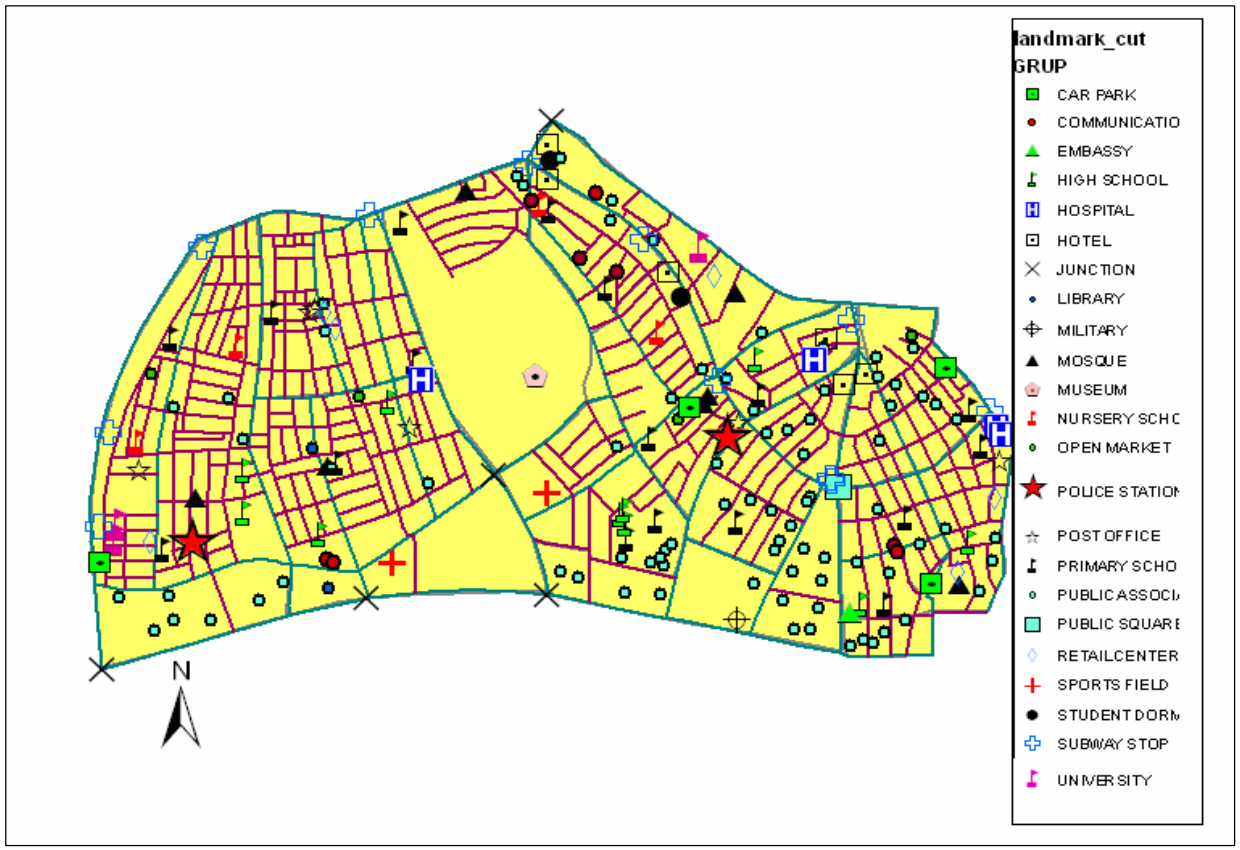



Figure 3.4.Land marks of the study area.

There are five different types of crime distributed over the area. Number of each type observed is illustrated in Figure 3.5, where the burglary has the highest number of incidents. All the incidents per crime type are mapped to demonstrate the spatial distribution in the study area (Figure 3.6).

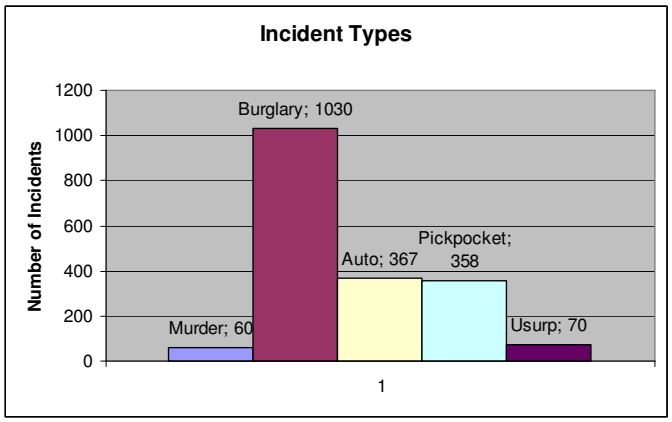



Figure 3.5.Number of incidents with respect to crime types in the study area.

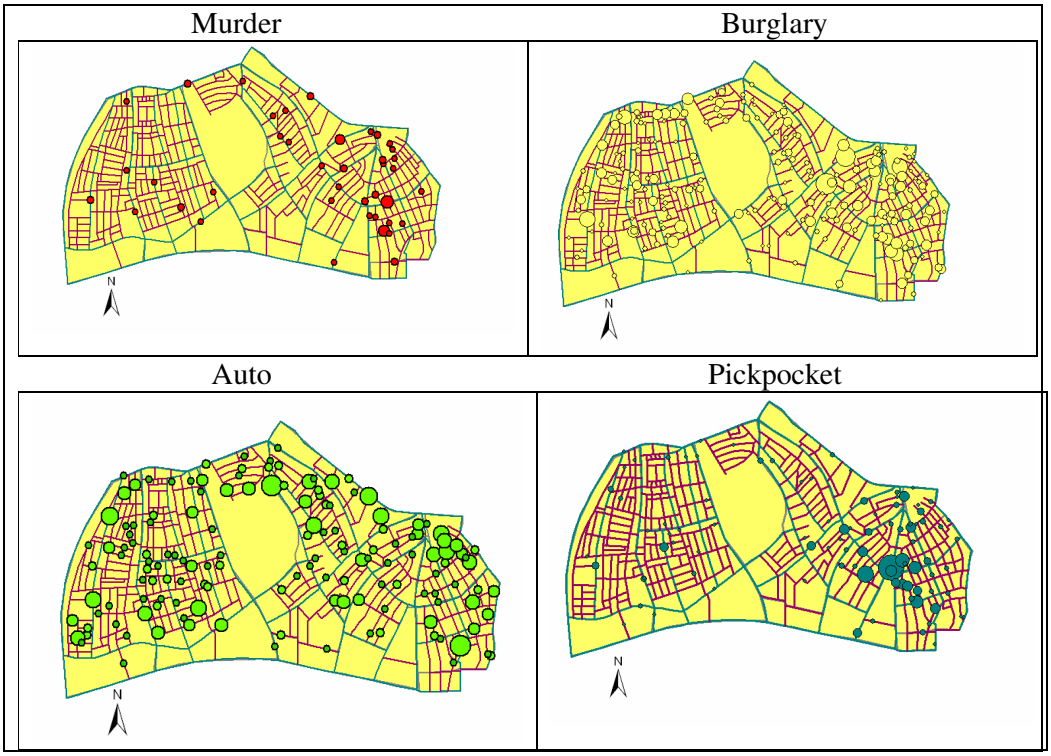



Figure 3.6.Spatial distribution of crime types in the study area.

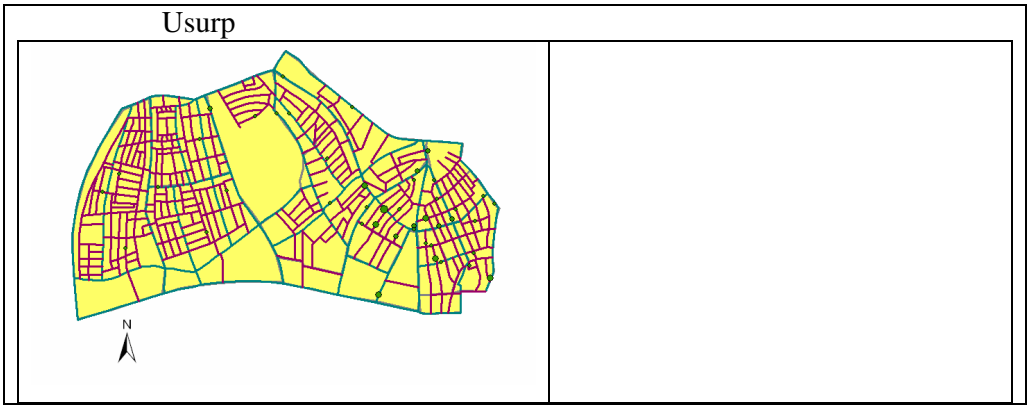


Figure 3.6.Spatial distribution of crime types in the study area (cont'd).

Regarding to maps (Figure 3.6), the distribution of burglary and auto related crime incidents are more randomly than the others. Auto related crime incidents are generally observed near the boundaries of the study area. Also, the number of burglary incidents near Atatürk Avenue is significant.

Murder is occurred mostly at Merkez Çankaya region, while there is no usurp incidents recorded in Bahçelievler. Pickpocketing is differing the other crime types in that although the number of incidents for auto related crimes and pickpocketing are similar, the distribution of incidents are totally different. Most of the pickpocketing incidents are located in Kızılay Square which depicted in Figure 3.3.

As the crime prediction model is based on weekday values, hence crime incidents for each day are mapped with the number of incidents per crime type per day (Figure 3.7).

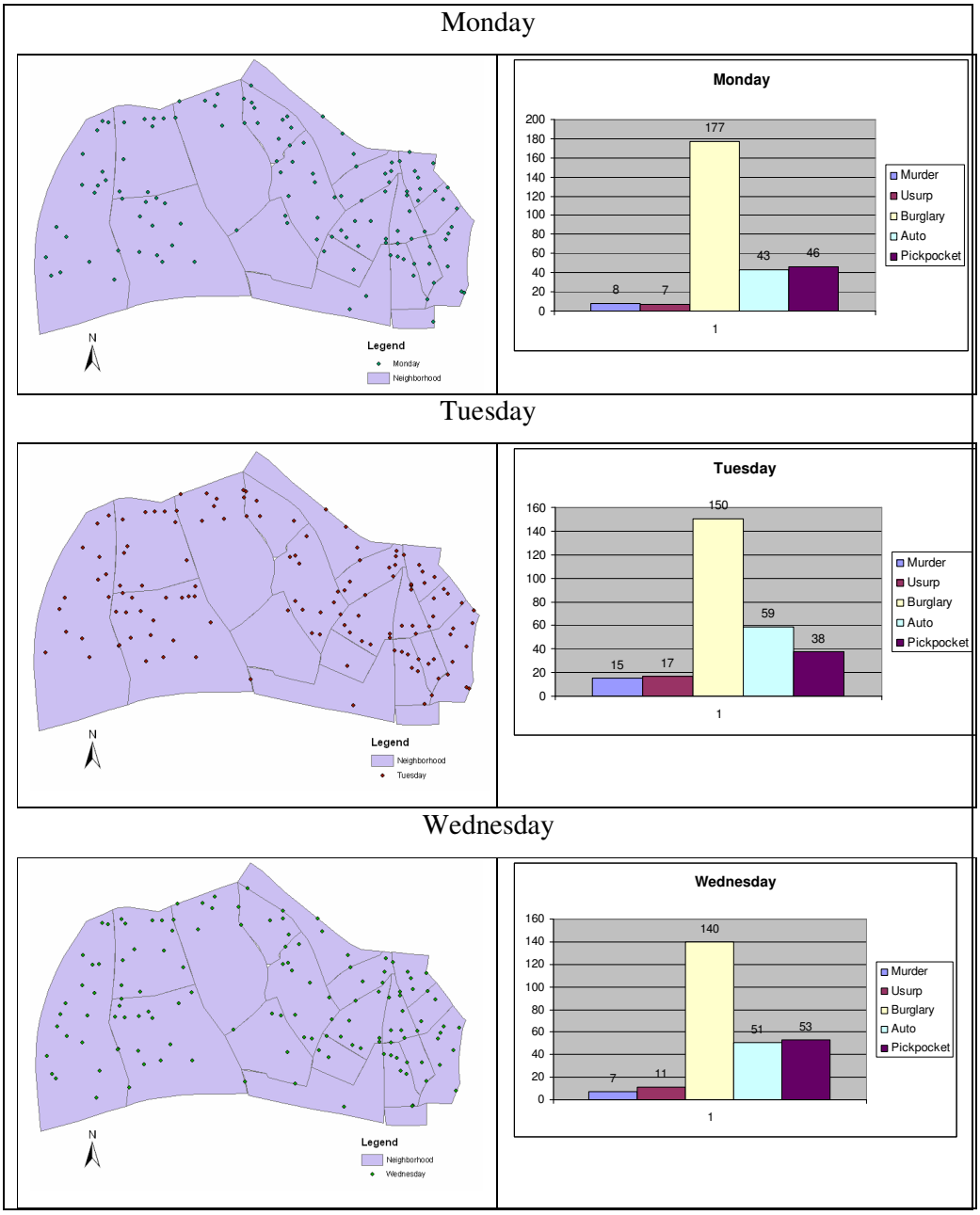



Figure 3.7.Spatial distribution of crime incidents per weekday and graphs of number of incidents per crime type per day.

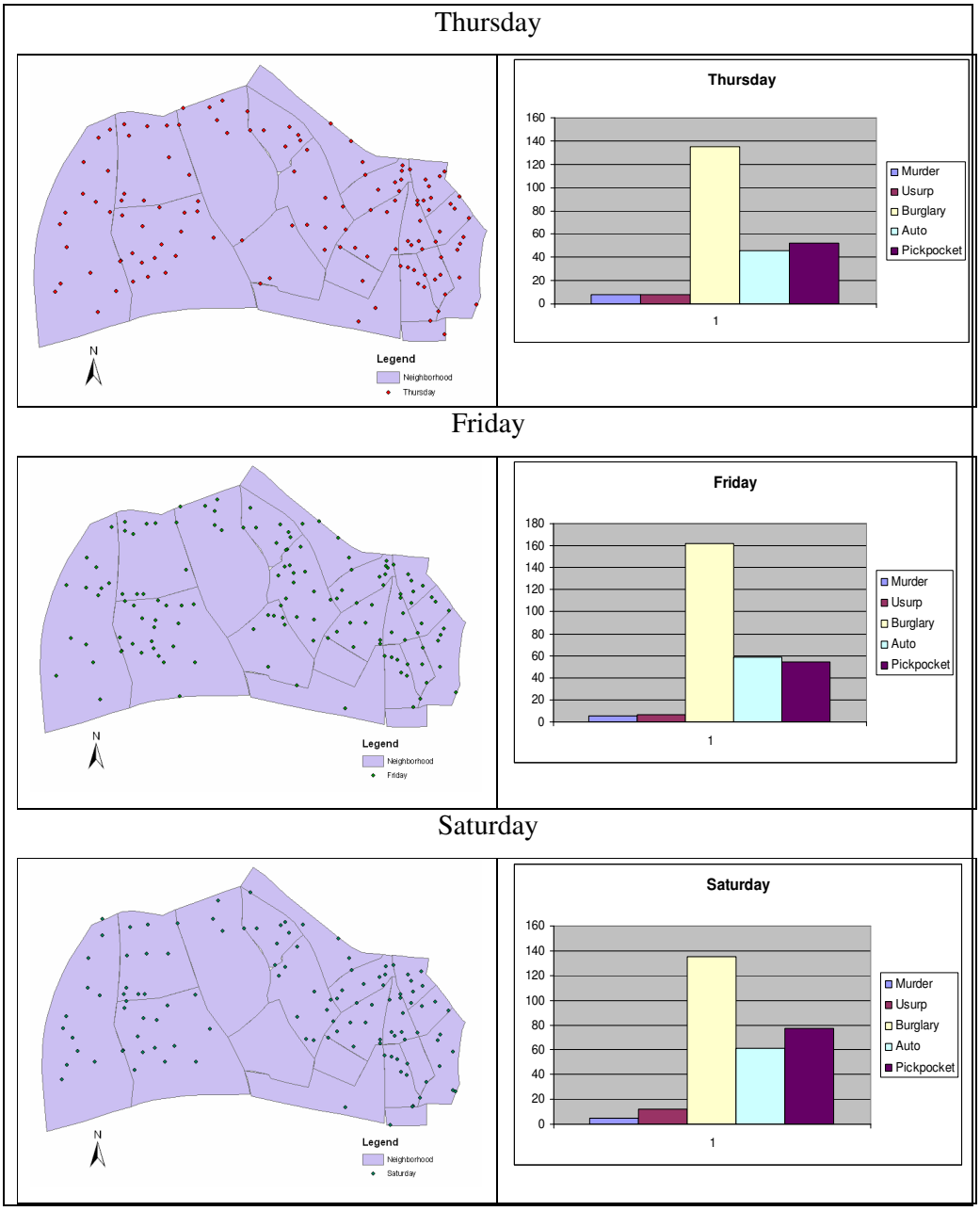



Figure 3.7.Spatial distribution of crime incidents per weekday and graphs of number of incidents per crime type per day (cont'd).

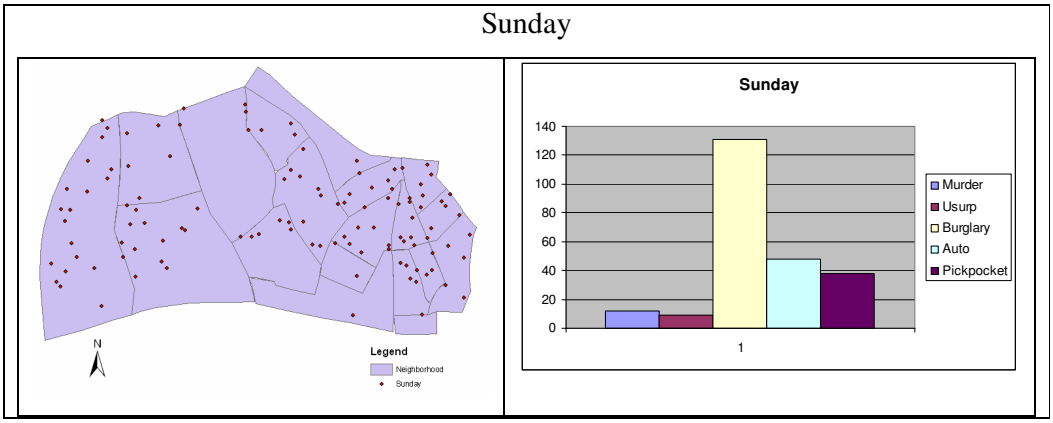



Figure 3.7.Spatial distribution of crime incidents per weekday and graphs of number of incidents per crime type per day (cont'd).

According to the Figure 3.7, distribution of crime incidents per each weekday are not significantly different when depicted with single symbol representation. However, there are deviations in the numbers and the orientation of the incidents for each day which is necessary to apply the proposed spatio-temporal crime prediction model. For example, number of pickpocketing at Saturday is 66% higher than the other days at average (Figure 3.7). This situation can be explained as Saturday is holiday so people prefer making shopping at Saturdays which gives many opportunities for offenders to commit crime. However, the number of crime types indicates similarity in terms of the highest and the lowest occurrence of the crime incidents according to crime types. As illustrated in Figure 3.7, the highest number of crime incidents occurred in the area is burglary, whereas usurp and murder are the least recorded crime types

### 3.3. Methodology of the study

The spatio-temporal crime prediction model in this study is adapted from model generated by Al Madfai et al., (2006). There are significant differences in two models as the number of crime incidents in Bahçelievler and Merkez Çankaya police precincts is nearly 10% of the number of data used in the adapted study. Differences and similarities are explained in methodology part. The methodology of the study is outlined in Figure 3.8.

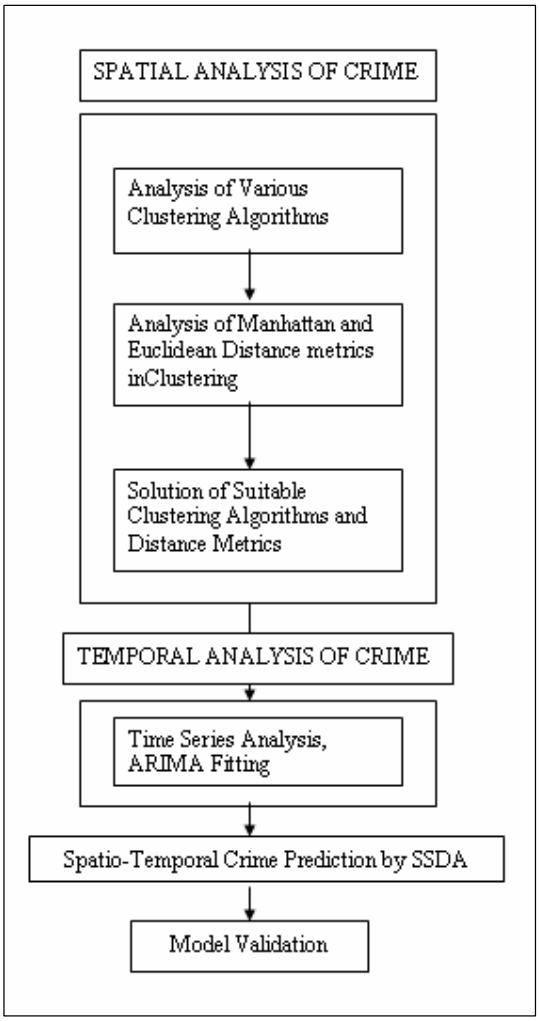



Figure 3.8.Methodology of the Study

The methodology starts with analysis of various clustering algorithms. These algorithms are K-means, Nnh Hierarchical, STAC, ISODATA, fuzzy, and GAM. Clustering algorithms are analyzed and compared with each other to get the most convenient algorithm to be involved in spatio-temporal crime prediction model. However, Al Madfai et al., (2006) do not consider different clustering algorithms in their study and apply STAC clustering for the crime prediction model.

Then different from the original work, different distance metrics, Manhattan and Euclidean are considered in the scope of the study. Both distance metrics are

applied to seek the difference between the orientations of the clusters. Distance metrics are used with the selected clustering algorithm. The reason behind this consideration is to compare both distance metrics in terms of clustering and predictive performance of the spatio-temporal crime prediction model. After the spatial analysis of crime in two paragraphs, the next step is to explain the temporal analysis of crime in the crime prediction model.

Al Madfai et al., (2006) use hierarchical profiling approach for temporal analysis of crime which was generated by their own. Approach has two stages to generate the forecasting model. First, special days of a year such as Noel, Halloween day are defined. Those are the days when the number of crime incidents deviate from the normal more than the average for their study. Then, selected days are profiled and modeled for the deterministic part of the study. The stochastic part is modeled by Box-Jenkins ARIMA model. After generating time series, forecasted values are obtained as a result for the three year data (Al Madfai et al., 2006). However, in our study with the data on hand, applying hierarchical profiling approach is not suitable as one year data is not adequate to observe the specific daily fluctuations in a year. In spite of hierarchical profiling approach, time series analysis with ARIMA fitting is used to model the data in this thesis. As a result forecasted values of one year data is obtained.

At last spatio-temporal crime prediction model is completed by disaggregating the forecasted values to the clusters. Clusters are daily clusters generated with STAC and model is validated with calculating spatio-temporal mean root square error for the entire model and for each cluster. Al Madfai et al. (2006) also calculated the error terms to validate the model.

An adequate spatio-temporal crime prediction model has several indicators. The first and the most important one is verification of the results by statistical testing. During the implementation process and also the results should be verified and validated. The second indicator when generating a crime prediction model is the orientation of the clusters and meaningful predictions for each cluster. To

understand the validity of the result by this way, the social, economical, and physical properties of the area should be investigated. In this thesis all these indicators are considered to prove the validity of the model.

### 3.3.1. Spatial analysis of crime

#### 3.3.1.1. Cluster Analyses

The classification of objects into different groups sharing the same characteristics is termed as clustering. Clustering is a common technique for data mining, image analysis, biology and machine learning. Techniques which search for separating data in to convenient groups or clusters are termed as clustering analysis (Everitt, 1974).

Ball (1966) listed seven possible uses of cluster analysis techniques:

- Finding a true typology,
- Model fitting,
- Prediction based on groups,
- Hypothesis testing,
- Data exploration,
- Hypothesis generating,
- Data reduction.

Cluster analysis techniques roughly classified into five types as follows (Everitt, 1974):

1. Hierarchical techniques: The classes themselves are classified into groups, the process being repeated iteratively resulting forming a tree structure.
2. Optimization-partitioning techniques: The classes are formed by optimizing by clustering criteria which form a partition of the set of entities.

3. Density or mode-seeking techniques: The clusters are formed considering the dense concentration of entities.
4. Clumping techniques: In this technique, overlapping of classes or clumps allowed.
5. Other methods not falling into the other categories.

Cluster analyses can provide significant insight into identifying crime patterns. The technique in crime phenomena is generally used for hot spot detection. Hot spots in crime analysis are one of the major explanations of criminal activities and spatial trends. Clustering is one of the reliable and objective methods to identify hot spots. However, the application of cluster analysis for hot spot detection has some problems. Researchers generally confused of deciding the appropriate clustering model. One of the reasons is the spatial nature of hot spots that if the clustering models coordinated accurately with geography. Another reason is the difficulty of determining the number of appropriate cluster which is not defined clearly in literature. Generally analysts use both visualization and statistical methods to be at the safe side of choosing the right number. Number of clusters can be defined according to the purpose of usage by visualization techniques.(Grubesic, 2006).

Among the various types of clustering algorithms; in this thesis K-means, Nnh hierarchical, STAC, Fuzzy, ISODATA are selected to be used. K-means, fuzzy, ISODATA are included in optimization-partitioning techniques, and Nnh hierarchical is a hierarchical clustering analysis technique. STAC and GAM are two clustering algorithms generated for specific usage purposes. Spatio-temporal analysis of crime was invented by crime analysts to detect crime hot clusters. Also, the algorithm of geographical analysis machine is generated according to the underlying population.

#### 3.3.1.1.1. K-means clustering

The K-means clustering method is a non-hierarchical clustering approach where data are divided into K groups. User is flexible to decide on the number K. The aim of this technique is to create K number of clusters so that the within group sum of squares are minimized. As iterating all the possible observations is enormous, the algorithm finds a local optimum. To reach the optima, algorithm is repeated several times and the best positioning K centers are found. Then the remaining observations to the nearest cluster to minimize the squared distance (Levine, 2002), are assigned. The formula of the algorithm is:

$$MinV = \sum_{i=1} \sum_{k} a_i d_{ik}^2 y_{ik} \ , \qquad \text{Eq. (3.1)}$$

Where,

i = index of observations;

$a_i$ = attribute weight of observation I;

k = index of clusters;

$d_{ik}$ = distance between observation I and cluster k;

$y_{ik}$ = binary value if observation I is assigned in cluster k;

The constraints of the model are assigning each observation to a cluster group and limiting decision variables to be an integer. In k-means clustering each observation should be assigned to only one group and all the observations have to be included by clusters. Taking squared Euclidean distance as a distance measure is important because the model is becoming non-linear and can be solved by appropriate heuristic approaches. Distance squared function gives too much importance to spatial outliers and the resulting clusters may subject to skewness (Grubesic, 2006).

In CrimeStat software the routine starts with an initial guess about the K locations. Initial guess is made by software randomly selecting initial seed locations. It is an

iterative procedure determining the initial seed and assigning the observations to the initial seed and then recalculates the center of the cluster and assigns observations again where the procedure stops (Levine, 2002).

K-means clustering routine outputs clusters graphically as either ellipses or convex hulls. For the standard deviational ellipses 1X, 1.5 X and 2X are the options to select. Here X represents the standard deviation from normal. The level of standard deviation is chosen by the user. Where, by standard deviational ellipses some of the data is abstracted, convex hull represents the cluster by drawing a bounding polygon outside the data (Levine, 2002).

#### 3.3.1.1.2.Hierarchical Clustering

Levine (2002) stated that hierarchical clustering is a clustering method which groups observations on the basis of defined criteria. Criteria is related with type of distance between observations. The clustering is repeated until all the observations are clustered according to the selected criteria and order. Based on the criterion selected, hierarchical clustering has several options called nearest neighbor, farthest neighbor, the centroid, median clusters, group averages and minimum error (Levine, 2002)

There are two basic hierarchical clustering methods implemented, agglomerative and divisive. Both of these methods are considering dissimilarities between the observations. As the geography is concerned in spatial analysis, dissimilarity measure to be taken into account should be distance in analysis of crime.

The general algorithm of hierarchical clustering is outlined as follows where the distance between two observations i and j is represented by $d_{ij}$ and cluster i contains $n_i$ objects. Let D be the remaining $d_{ij}$'s. Suppose there are N objects to be clustered.

- Find the smallest element remaining in **D.**

- Merge clusters i and j into a single new cluster, k.
- Calculate a new set of distances $d_{km}$ using the following distance formula:

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma \left| d_{im} - d_{jm} \right| \qquad \text{Eq. (3.2)}$$

Where;

m represents any cluster other than k. These new distances replace $d_{im}$ and $d_{jm}$ in **D**. Also let, $n_k = n_i + n_j$

Note that the algorithms available represent choices for $\alpha_i$, $\alpha_j$, $\beta$ and $\gamma$.

- Repeat steps 1 - 3 until **D** contains a single group made up off all objects. This will require *N-1* iterations (Web8).

Controlling the size of grouping either with threshold distance and minimum number of observations in clusters give chance to identify dense small geographic environments is one of the advantages of hierarchical clustering. Hierarchical clustering can calculate first, second and higher order clusters to identify hot spots in different levels. In addition, each of the levels implies different policing strategies. First order clusters indicate small neighborhoods where the second order clusters indicates patrol areas. Thus, the hierarchical technique allows different security strategies to be adopted and provides a coherent way of approaching these communities (Levine, 2002)

There are also some disadvantages to this clustering method. The method only considers points but not weights or attributes. Another is, when the method makes a mistake at a one stage, it is impossible to turn back and revise the results. At last, the method is subjective and there is no solid theoretical background behind it (Grubesic, 2006).

Nearest neighbor (Nnh) hierarchical clustering algorithm is a type of hierarchical clustering algorithm. To adjust the algorithm given in hierarchical clustering according to nearest neighbor distance between algorithms, following values are given to coefficients in the model:

The distance between two groups is defined as the distance between their two closest members. The algorithm is the same but the coefficients change, where;

$\alpha_i$ and $\alpha_j = 0.5$,

$\beta = 0$ and $\gamma = -0.5$.

The CrimeStat Nnh routine, which is also used in the thesis, uses a method defining a threshold distance and minimum number of points to be included in the cluster. Two choices of defining threshold distance are random nearest neighbors with fixed distance. After distance selection procedure, number of minimum points in a cluster should be determined. This criterion is used to reduce the number of very small clusters. The default is 10 but user can change the number according to the purpose of the study (Levine, 2002).

The output of CrimeStat gives standard deviational ellipses and convex hulls of clustered data. There are advantages and disadvantages of each approach. The convex hull has the advantage of being a polygon that covers exactly all the cluster. The ellipse, on the other hand, is more general and will usually be more stable from year to year. It usually looks better on a map and users are able to understand it better. The biggest disadvantage of an ellipse is that it forces a certain shape on the data, whether there are incidents in every part of it or not (Levine, 2002).

#### 3.3.1.1.3.Fuzzy Clustering

Fuzzy clustering is a generalization of partition clustering methods (such as k-means and medoid) by allowing an observation to be classified into one or more clusters. Observations have probability of belonging to one or more cluster in

fuzzy clustering. While in regular clustering one observation has to be a member only one cluster. In fuzzy clustering, the membership is spread among clusters having a probability of membership value ranging from 0 to 1. Not forcing every object to a specific cluster is really an advantage of this method especially where the spatial outliers exists. However, there are much more information on hand to be handled (Web8).

Kaufmann (1990) describes the fuzzy algorithm as:

$$MinZ = \sum_{k} \frac{\sum_{i}\sum_{j} u_{ik}^{2} u_{jk}^{2} d_{ij}}{2\sum_{j} u_{jk}^{2}} \qquad \text{Eq. (3.3)}$$

Where;

i,j = index of observations,
k = index of clusters,
$d_{ij}$ = distance between objects i and j,
p = number of cluster groups,
$u_{ik}$ = fractional membership of observation i in cluster k.

Minimizing the objective function subject to constraints requires for each observation to have a constant total membership distributed over the different clusters. Also, memberships can not be negative. The objective function forces to minimize the total dispersion.

Fuzzy clustering method is a part of partitioning clustering methods differing when fractional membership of observations (u) are taking values less than 1. The goal of the fuzzy clustering method is to minimize the aggregate distance between observations and clusters including the dissimilarity of not weighting observations (Grubesic, 2006).

#### 3.3.1.1.4.Geographical analysis machine

Geographical analysis machine is an automated exploratory spatial data analysis that is detecting localized clustering while; other clustering methods are concentrated on global measures of the pattern. The results are invariant in terms of study region boundary and visualized by cartographic measure bringing the user to understand the clusters easily. One other advantage is that end-users who do not have degree in statistics are able to understand the resulting of the analyses. However, it is computationally difficult and takes long run times for obtaining the results (Openshaw, 1998).

The GAM algorithm involves the following steps:

- Read in X, Y data for population at risk and a variable of interest from a GIS.
- Identify the rectangle containing the data, identify starting circle radius, and degree of circle overlap.
- Generate a grid covering this rectangular study region so that circles of radius r overlap by the desired amount.
- For each grid-intersection generate a circle of radius.
- Retrieve two counts for this circle.
- Apply some significance test procedure.
- Keep the result if significant.
- Repeat steps 5 to 7 until all circles have been processed.
- Increase circle radius and return to step 3 else go to step 10.
- Create smoothed density surface of excess incidence for the significant. circles using a kernel smoothing procedure.
- Map this surface as it peaks that suggest where the accumulated evidence of clustering is likely to be greatest (Openshaw, 1998).

#### 3.3.1.1.5. ISODATA

The ISODATA is an iterative self-organizing way of performing clustering. The minimum Euclidean distance is used as the index to assign each candidate cell to a cluster, and the self-organizing method is used as the optimal solution to optimize the cluster classes (Web9). ISODATA clustering is generally used for image processing applications and is not commonly used in crime related analysis. However, in order to understand the suitability of this method in crime analyses, it is used in this thesis.

#### 3.3.1.1.6. STAC

The spatial and temporal analysis of crime (STAC) was developed by the Illinois Justice Information Authority as a clustering algorithm in which the number of points is counted in a circle laid over a grid. The STAC routine in Crime Stat identifies densest clusters by demonstrating either standard deviational ellipses or convex hulls (Levine, 2002).

STAC is adopted as the other clustering routines; however, it differs in the process that the overlapping clusters are combined into larger clusters until there are no overlapping clusters. Both partitioning and hierarchical clustering routines are included in the STAC. Search circles and aggregating small clusters into larger ones are some properties of the other clustering routines. Another advantage of STAC is that it is not constrained by artificial or political boundaries, such as police precincts or census tracks (Levine, 2002).
STAC in CrimeSTAT follows the procedure:

1. Overlaying of a 20 x 20 grid structure on the plane defined by the area boundary (defined by the user).

2. Placement of a circle on every node of the grid, with a radius equal to 1.414 times the specified search radius. Thus, the circles overlap.

3. Counting the number of points falling within each circle, and ranking the circles in descending order.

4. Recording all circles with at least two data points along with the number of points within each circle for a maximum of 25 circles. The X and Y coordinates of any node with at least two incidents within the search radius are recorded, along with the number of data points found for each node.

5. Ranking circles according to the number of points and selection of the top 25 search areas.

6. Combining points within circles if a point belongs to two different circles. This process is repeated until there are no over lapping circles. This routine avoids the problem of data points belonging to more than one cluster, and the additional problem of different cluster arrangements being possible with the same points.

7. Calculating the best fitting standard deviation ellipse or convex hulls by using data in clusters (Levine, 2002).

### 3.3.2. Temporal analysis of crime

#### 3.3.2.1. Univariate Box-Jenkins (ARIMA) Forecasting

Data in business, economics, engineering, crime and other sciences are often collected in the form of time series. A time series is a set of values observed sequentially at regular intervals of time such as weekly traffic volume, daily crime rates, and monthly milk consumption. The main objective of the time series analysis are to understand the underlying and time-dependent structure of the single series-univariate series and to figure out the leading, legging and feedback relationships (Pena et al., 2001).

Univariate Box-Jenkins is a time series modeling process which describes a single series as a function of its own past values. To find an appropriate equation that reduces a time series with underlying structure to white noise is the aim of the Box-Jenkins process. The reason of the popularity of the Box-Jenkins modeling process is that it uses the data itself to determine appropriate model form, whereas many other time series modeling methods use given model as an assumed model for priori. For a given time series to find the best model for that series, the model form should be determined carefully (Web8).

Box-Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. The method is suitable for time series that have at least 50 observations.

The model is generally referred to as an ARIMA (p,d,q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model, respectively. When d = 0, the model is turned to be an ARMA (p,q) model.

Autoregressive integrated moving average (ARIMA) modeling is formed from two parts: the self-deterministic part and the disturbance component. The self-deterministic part of the series should be forecastable from its own past by an autoregressive (AR) model. Each autoregressive factor is a polynomial of the form:

$(1 - \Phi_1 B_1 - \Phi_2 B_2 - \Phi_3 B_3 - \ldots\ldots\ldots\ldots\ldots \Phi p B p)$,

Where $\Phi_1 \ldots\ldots\ldots \Phi p$ are the parameter values of the polynomial, and B is the backshift operator. The values of the autoregressive factors ($\Phi_1 \ldots\ldots \Phi p$) need not all be nonzero. A zero parameter value indicates that the parameter is not included in the polynomial.

The disturbance component (the residuals from the autoregressive model) is modeled by a moving average (MA) model. Each moving average factor is a polynomial of the form:

$(1 - \theta_1 B_1 - \theta_2 B_2 - \theta_3 B_3 - \ldots\ldots\ldots\ldots \theta qBq)$,

Where $\theta_1 \ldots \theta q$ are the parameter values of the polynomial and B is the backshift operator. The values of the $\theta p \ldots \theta q$ need not all are nonzero. A zero parameter value indicates that the parameter is not included in the polynomial.

The backshift operator is a special notation used to simplify the representation of lag values. $B_J X_t$ is defined to be $X_{t-J}$. So, $(B_1)X_t = X_{t-1}$ which means a 1 period lag of X (Brockwell, 1996).

Hence the ARIMA(p, d, q) model is:

$$\Phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t , \qquad \text{Eq. (3.4)}$$

Where $\varepsilon_t$ is an error term generally assumed to be independent, identically distributed samples from a normal distribution, d is a positive integer that controls the level of differencing (Pena et al., 2001).

The Box-Jenkins method refers the procedure involves making successive approximations through three stages: identification, estimation, and diagnostic checking: (Web8).

**Identification stage:** In order to decide a tentative model form, time series requires examining the identification phase. The stage controls if the series is sufficiently stationary (free of trend and seasonality) and estimate the levels of the p, d and q parameters. The first part of the identification stage is to be ensuring that the time series is stationary. In a stationary series, the observations fluctuate about a fixed mean level with a constant variance over the observational period.

There are two types of non-stationarity in time series. To create a mean stationary series, differencing in sufficient level is applied and to adjust a variance stationary series, correct power transformation is applied. Unit root tests are applied to the model to test the stationarity of the model. Phillips-Perron and Dickey-Fuller tests are types of unit root tests.

Autocorrelations and partial autocorrelations are used extensively in the identification phase of the time series analysis. When plotted, they become the correlogram which visualizes the estimates of autocorrelations.

The correlation between $X_t$ and $X_{t+k}$ is called the $k^{th}$ order autocorrelation of X. The sample estimate of this autocorrelation, called *r*, is calculated using the formula:

$$r_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \overline{X}\right)\left(X_{i+k} - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad \text{Eq.(3.5)}$$

Where,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{Eq. (3.6)}$$

The $k^{th}$ order partial autocorrelation of X is the partial correlation between $X_t$ and $X_{t+k}$ , where the influence of $X_{t+1}, X_{t+2}, \ldots\ldots\ldots X_{t+k-1}$ have been removed (Web8).

Autocorrelation plots assist to understand what type of differencing is needed to reach mean stationary series. The decreasing pattern of the autocorrelation graph indicates the level of differencing and provides criteria for specification of p and q. The need for a power transformation can be ascertained by examining plots of both the original series and the transformed series (Web7). Autocorrelation

function is always 1 at the first lag. The series needs differencing when the function decays slowly without reaching zero. Seasonal differencing is determined by the number of time periods between the relatively high autocorrelations (Figure 3.4.1) (Web8).

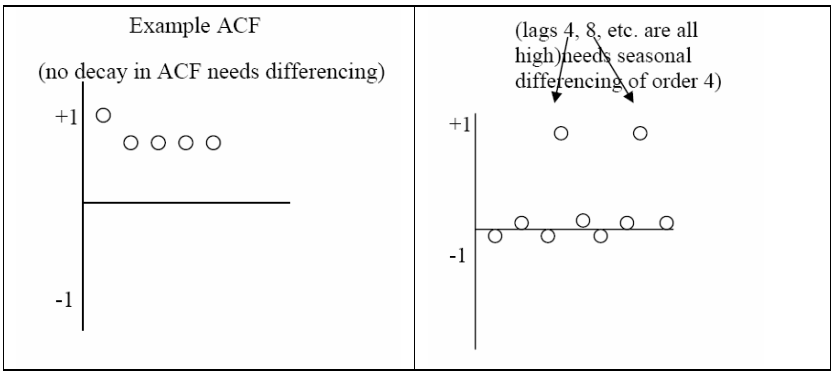



Figure 3.9.Examples of Autocorrelation functions of time series need differencing.

If the time series is non-stationary with respect to its variance, then the variance of the time series can be stabilized by using power transformations (Bowerman and O'Connell, 1993). Box-Cox is a one way of power transformations that can be represented as:

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda} for \lambda \neq 0 \end{cases} \qquad \text{Eq.(3.7)}$$

$$Y_t^{(\lambda)} = \log Y_t for \lambda = 0 \qquad \text{Eq.(3.8)}$$

Where, $Y_t^{(\lambda)}$ is a positive random variable (Hamilton, 1994).

**Estimation stage:** The autoregressive and moving average parameters of the selected model are estimated.

**Diagnostic Checking**: After the AR and MA parameters are estimated, the model is checked whether the model fits the historical data adequately. The differences (residuals) between the original and forecasted data are judged to be sufficiently small or random. If residuals are not satisfactory, the model is improved to enhance the predictability of the model.

### 3.3.2.2. Simple Spatial Disaggregation Approach

Simple spatial disaggregation approach (SSDA) is namely a spatio-temporal forecasting technique. The approach aims to produce cluster forecasts that give minimized forecast errors. Simple spatial disaggregation approach explores and identifies week day specific clusters, different from each other. In simple spatial disaggregation approach, it is assumed that distribution of crime incidents behave same on a weekday. This is not the real case but to obtain a reliable and straightforward method, deviations are ignored (Al Madfai et al., 2006).

Using the time series model, disaggregation into clusters are made according to the equation (Al Madfai et al., 2006):

$$\sum_{j}^{m_t} O_{tj} = y_{t-1}(1) \qquad \text{Eq.(3.9)}$$

Where;

$O_{tj}$ = Forecast of crime cluster j on day t.

$y_{t-1}(1)$ = One-step-ahead crime forecast for day t.

$m_t$ = total number of identified clusters at day t.

Based on equation (3.9) assigning forecasted values to clusters is done according to the formulation (Al Madfai et al., 2006):

$$O_{tj} = B_{tj} * \frac{y_{t-1}(1)}{m_t}, \text{ with } \sum_{\forall j} B_{tj} = 1 \qquad \text{Eq. (3.10)}$$

Where;

$B_{tj}$ is the spatial forecast disaggregation (SFD) weights allocated to each cluster per day.

To calculate the spatial forecast disaggregation (SFD) weights, four methods are proposed by Al Madfai et.al. (2006): The naïve forecasting method, the ordinary least square method on number of incidences, the arithmetic mean and the ordinary least square method on percentages of crime.

The spatio-temporal forecast errors are calculated with spatio-temporal mean root square error. The formulation of STMRSE is;

$$STRMSE(\varepsilon) = \sqrt{\frac{1}{n}\sum_{t}^{n}\sum_{j}^{m_t}\frac{(Observed_{tj} - O_{tj})^2}{m_t}} \qquad \text{Eq. (3.11)}$$

Where n being the total number of days (Al-Madfai et al., 2006).

# CHAPTER 4

## GENERATION AND INTERPRETATION OF CLUSTERS AND COMPARISON OF DIFFERENT CLUSTERING METHODS IN THE STUDY AREA

In this chapter, the aim is to decide the suitable clustering technique to be used in spatio-temporal crime prediction model. With respect to this purpose, the main concern is to detect clusters based on different clustering approaches. Rattcliffe (2004) explained that a hot spot is an area with high crime density. Predicting crime through hot spotting is a new advance to police departments to make tactical, strategic and administrative policies and to get right prevention measures. Clustering is gaining importance as it is a reliable way of determining crime hot spots. However, there is a strong confusion regarding using a convenient clustering model to detect hot spots. Choosing an appropriate clustering model is not easy in terms of general clustering approaches, number of clusters and geographical fitting. Data and statistical analysis should be carefully determined and evaluated to take advantage of clustering analysis in pro-active policing.

In order to make comparison between the clustering models, different clustering methods are applied to the study area, Bahçelievler and Merkez Çankaya police precincts. Both hierarchical and non-hierarchical/partitioning approaches are considered. K-means, fuzzy and ISODATA clustering algorithms are applied as partitioning based clustering approaches. These methods are including optimization procedures to get final configurations. Hierarchical clustering algorithm represents the hierarchical approach while spatio-temporal analysis of crime and geographical analysis machine are generated specifically for cluster detection. CrimeStat 3.1, GAM\K tool and TNTmips 6.4 are run to carry out analysis and results are interpreted with ArcGIS 9.1 and TNTmips 6.4, GIS softwares.

### 4.1. Application of K-Means Clustering

"K" is the key part of K-means clustering, which represents the number of clusters. It is not easy to determine the number of clusters in K-means clustering. Freedom of defining the number of clusters can be an advantage or disadvantage according to the purpose of usage. If too many clusters are generated, there will be patterns that are not really exist and also, too few clusters mean poor differentiation of the observations.

In order to determine the number of clusters, different K values are examined to determine the best configuration. CrimeStat 3.1 and ArcGIS 9.1 are employed in order to apply K-means clustering to the crime incident data in the study area. After trial period, it is found sufficiently to demonstrate the clusters from 5 to 8. The two reasons of choosing these values are: visualization and total mean squared error, which is an indicator to assist the decision of the K value. Error value is so high before number 5 and does not change much after 8 (Table 4.1). Hence, different K values (5, 6, 7, 8) are implemented to indicate the effect of difference in "number of clusters". It should be taken into account that total mean squared errors with different K values in Table 4.1. indicate a higher reduction at 7 clusters.

Table 4.1.Total mean squared errors of k-means clustering with different K values.

| Number of Clusters | Total mean squared error |
|---|---|
| 4 | 0.11 |
| 5 | 0.070 |
| 6 | 0.076 |
| 7 | 0.060 |
| 8 | 0.055 |
| 9 | 0.053 |

Visualization of the results is another concern deciding the number of clusters. Two techniques are available in CrimeStat 3.1: standard deviational ellipses and

convex hulls. In standard deviational ellipses, it is optional to decide on the size of the ellipses which are 1X, 1.5X and 2X. Here X represents the amount of standard deviation in algorithm. 1X is generally preferred as the other options gave an exaggerated view of the underlying clusters. Both of the visualization methods are mapped above with convex hulls and standard deviational ellipses -1X size clusters. For K=6, standard deviational ellipses and convex hulls are overlaid to indicate the difference (Figure 4.1). In convex hulls, "each object must belong to at least one group" constraint can be obviously seen while in standard deviational ellipses some of the observations are abstracted.

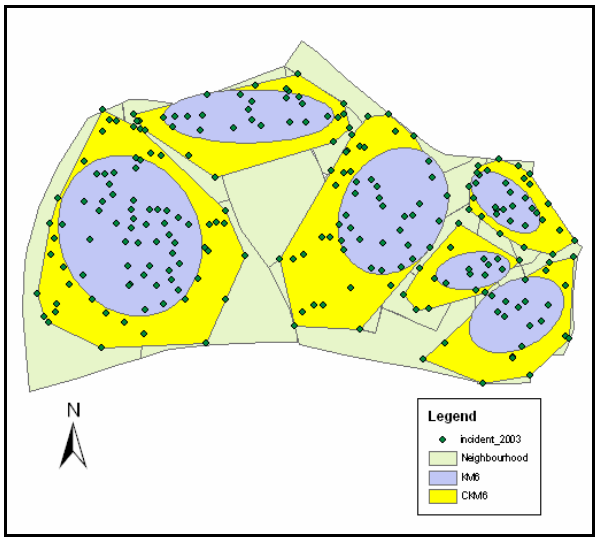



Figure 4.1.K-means clustering representations: standard deviational ellipses and convex hulls for K=6

As seen clearly, for all K values most of the partitioning took place at east part of the study area, Merkez Çankaya. Analyzing the results according to the regions; in fact it is observed that there are no significant differences for different K values in the clusters. Clusters are investigated at three main parts; Bahçelievler, Anıttepe-Beşevler and Merkez Çankaya. Up to the value of K, clusters are sub-divided in these areas. To be more informative, all maps are overlaid with land-uses (Figure 4.2).

For 5-means, 1st cluster includes Bahçelievler and Emek where Aşgabat Street (known as 7th street) is passing through. 2nd cluster is formed in Beşevler, while 3rd cluster contains parts of Strazburg Street and Necatibey Street and Çankaya Merkez police station. Also, Atatürk high school and Maltepe bazaar are two important crime appealing places included in that cluster.

The first three clusters are mostly located in residential areas, where some commercial areas exist especially at two sides of the main streets. 4rd cluster covers buildings of Ministry of Health and Sıhhiye Bazar. Last cluster contains Kızılay square, Meşrutiyet and part of Ziya Gökalp Street having a boundary with İzmir Street. The 4rd and 5th cluster is located in commercial areas like Kızılay square.

When the number of clusters increases to 6, first 4 clusters stay at the same location but last cluster is divided into two clusters. Kızılay square, İzmir and Mithatpaşa Streets are included in the 5th cluster and the remaining Meşrutiyet, Konur and Karanfil streets are covered by 6th cluster. In 7 cluster configuration, clusters in Merkez Çankaya do not change, while west part divided according to Emek and Bahçelievler neighbourhoods. 8 clusters represent the same places with the 7 cluster but the main difference is the 3rd road in Namık Kemal Street. The crime rates are high in that street but it is firstly represented by a cluster.

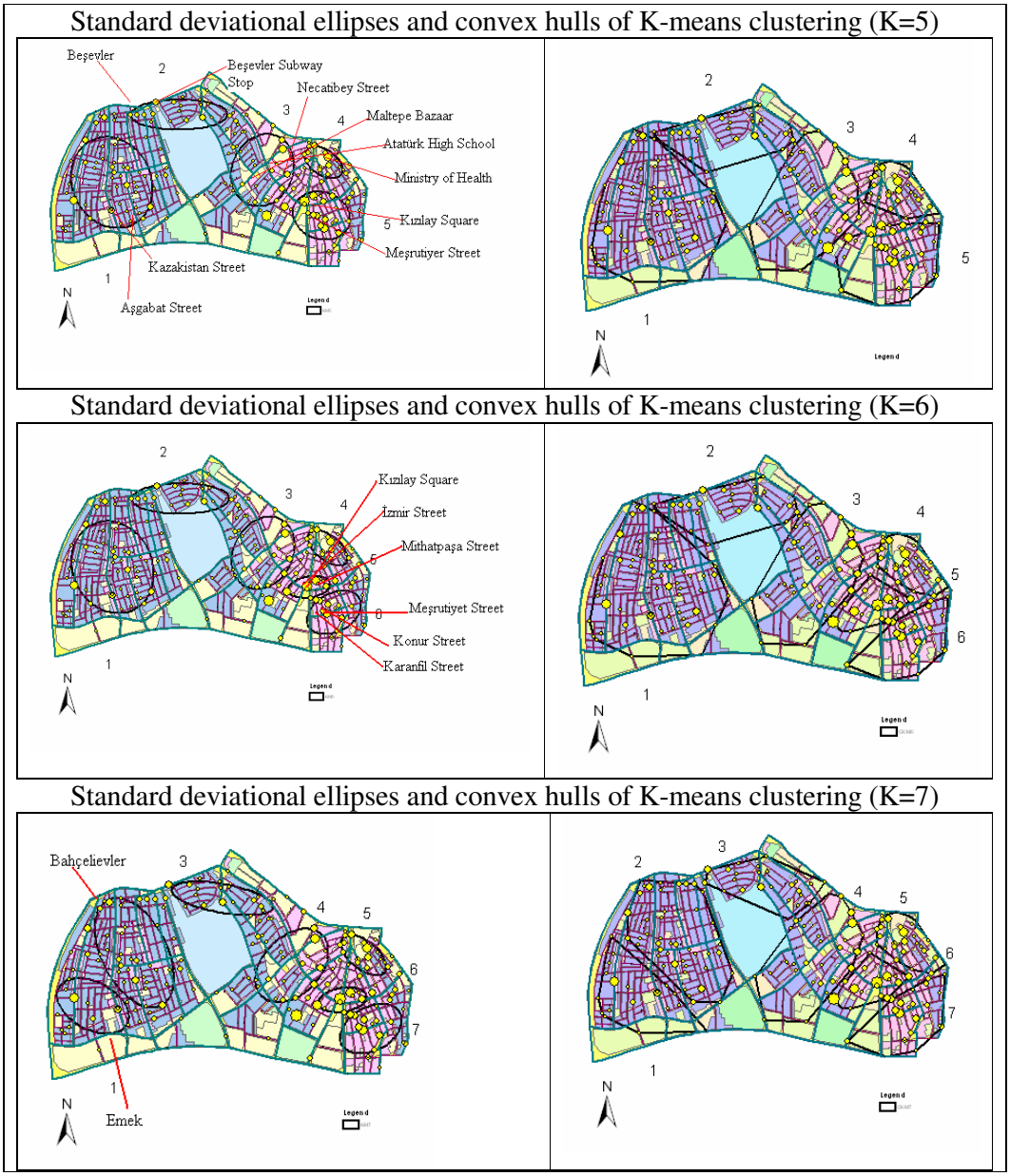



Figure 4.2.K-means clustering for the incidents

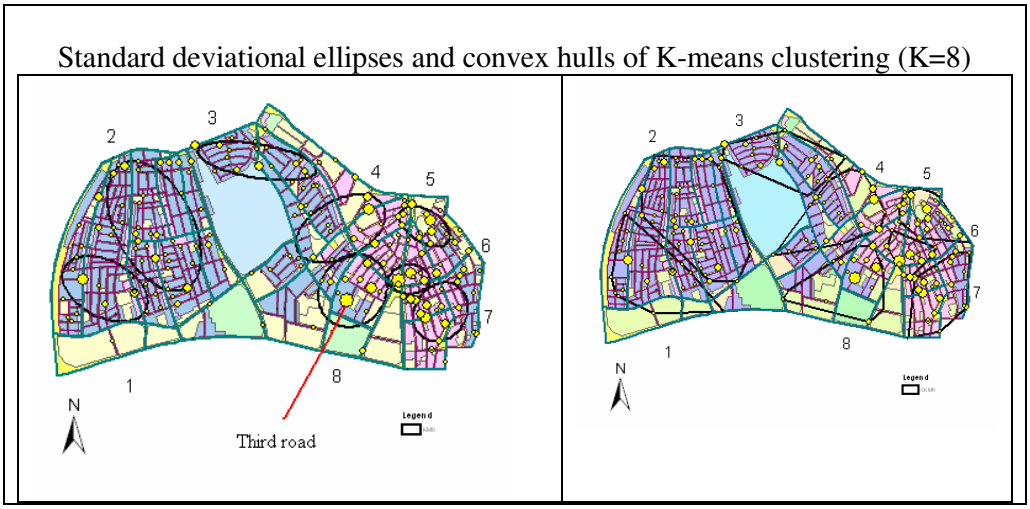


Figure 4.2.K-means clustering for the incidents (cont'd)

### 4.2. Application of Nnh Hierarchical Clustering

Nnh clustering is an agglomerative procedure, taking the observations individually and forms first order clusters based on a defined threshold distance and minimum number of observations in clusters. If two observations are nearer than the threshold value, a new cluster is generated. The second and higher order clusters are formed with the same manner until only one cluster is left or the threshold criteria fails. Two choices of defining threshold value are, random distance determined by the software itself and fixed distance defined by the user. When random distance option is selected, too many clusters are generated, so different fixed distance (300, 400, 500, 600 m.) options are tried to get the best configuration. Levine (2002) suggested to take the threshold value 0.5 miles or smaller to get feasible results. Also, after interpreting lots of "minimum number of points", 10 are selected visually. The resulting maps including first order clusters are shown in Figure 4.3 .

As stated in early chapters, hierarchical clustering approach does not necessarily cover all the observations in the area that can be seen in maps (Figure 4.3). The number of first order clusters is respectively high in hierarchical clustering as the

algorithm determines clusters based on geographical proximity. Analysis with fixed distances 300 and 400 meters confirm the algorithm and give small sized and high number of clusters. The results of this threshold values are valuable when a street or a specific area is being considered by a police or a crime analysts. The second and the higher order clusters provide different and more general view of the observations. One of the biggest advantages of Nnh hierarchical approach is to give opportunity to see different order of clusters at the same time and analyze the current situation with respect to the purpose.

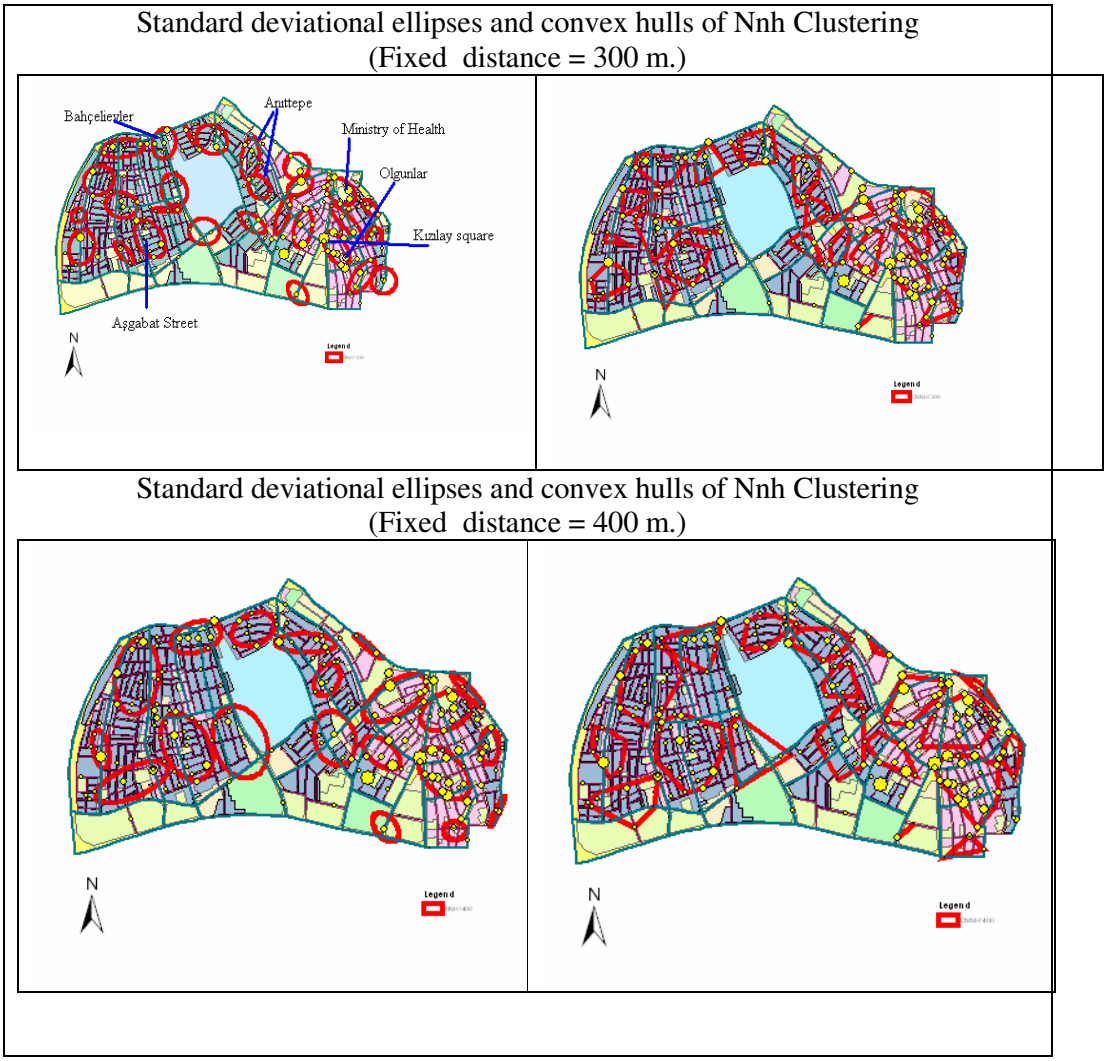



Figure 4.3.Nnh clustering for the incidents

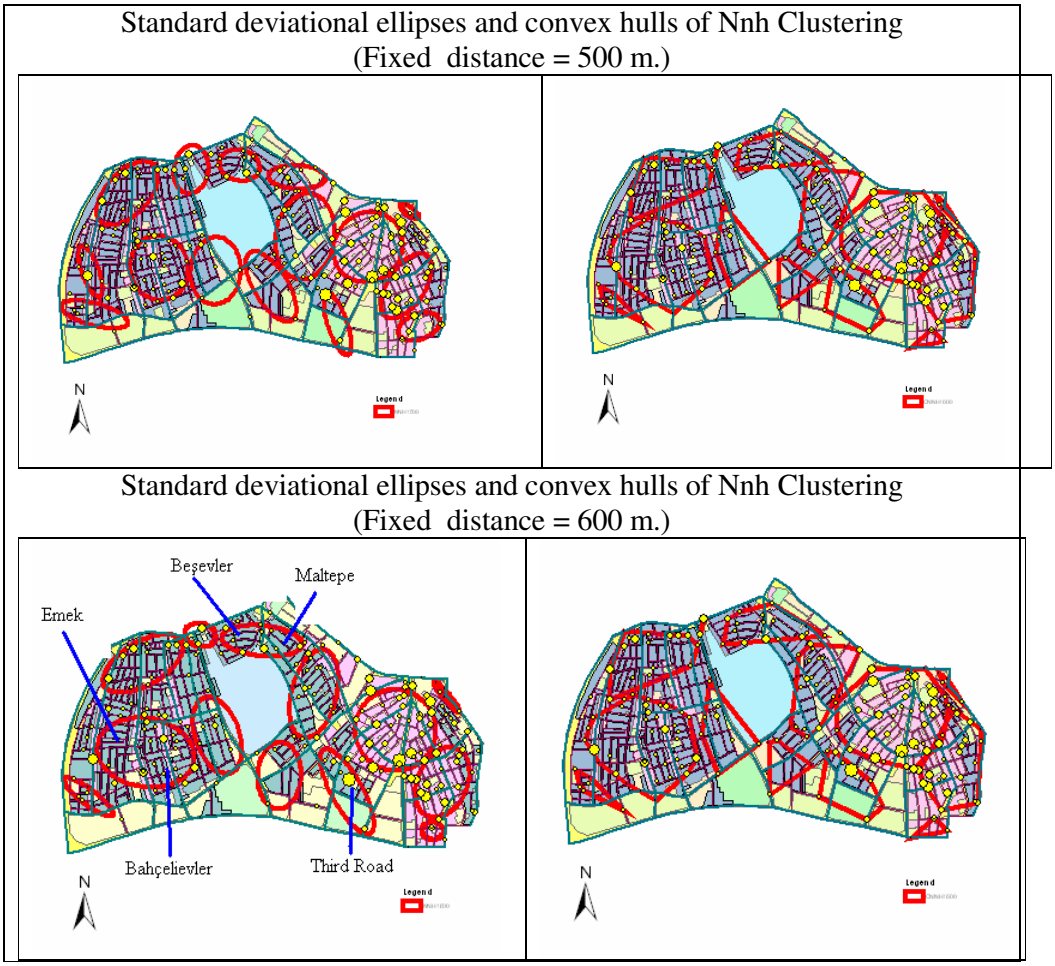


Figure 4.3..Nnh Clustering for the incidents (cont'd)

As explained in the previous clustering model, there are two visualization techniques are available in CrimeStat 3.1. In Figure (4.4), both interpretations are overlaid for Nnh clustering with fixed distance 600 m. The reason of choosing 600 m. is random to show the visualization techniques. In standard deviational ellipses representation and perception of the stakeholders are clearer than minimum bounding polygons. Minimum bounding polygons cover larger area than standard deviational ellipses as expected. 1X standard deviational ellipses consist of more than %50 of the observations having an area nearer to center.

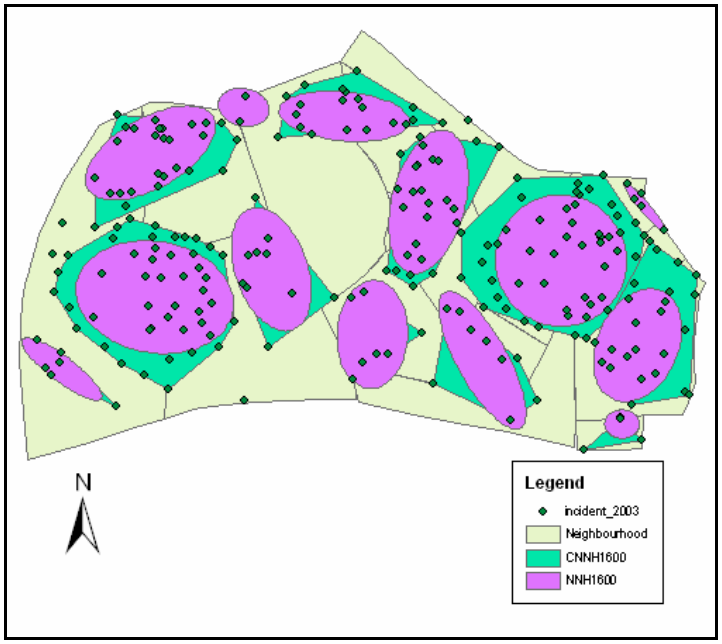


Figure 4.4. Nnh clustering representations: standard deviational ellipses and convex hulls for fixed distance 600 m.

In Nnh clustering lots of clusters generated, while most of them are located in Merkez Çankaya police precinct. The main underlying reason is that bigger part of the crime incidents took place in this part of the study area. Generally more incidents mean more clusters especially when nearest neighborhood distances are considered. The orientation and sizes are different according to the minimum number of points in clusters and distance between the incidents. For example, a coordinate where 10 incidents happened and no other incident within fixed distance is found, can be very small cluster like the clusters between Anıttepe and Bahçelievler (Figure 4.3).

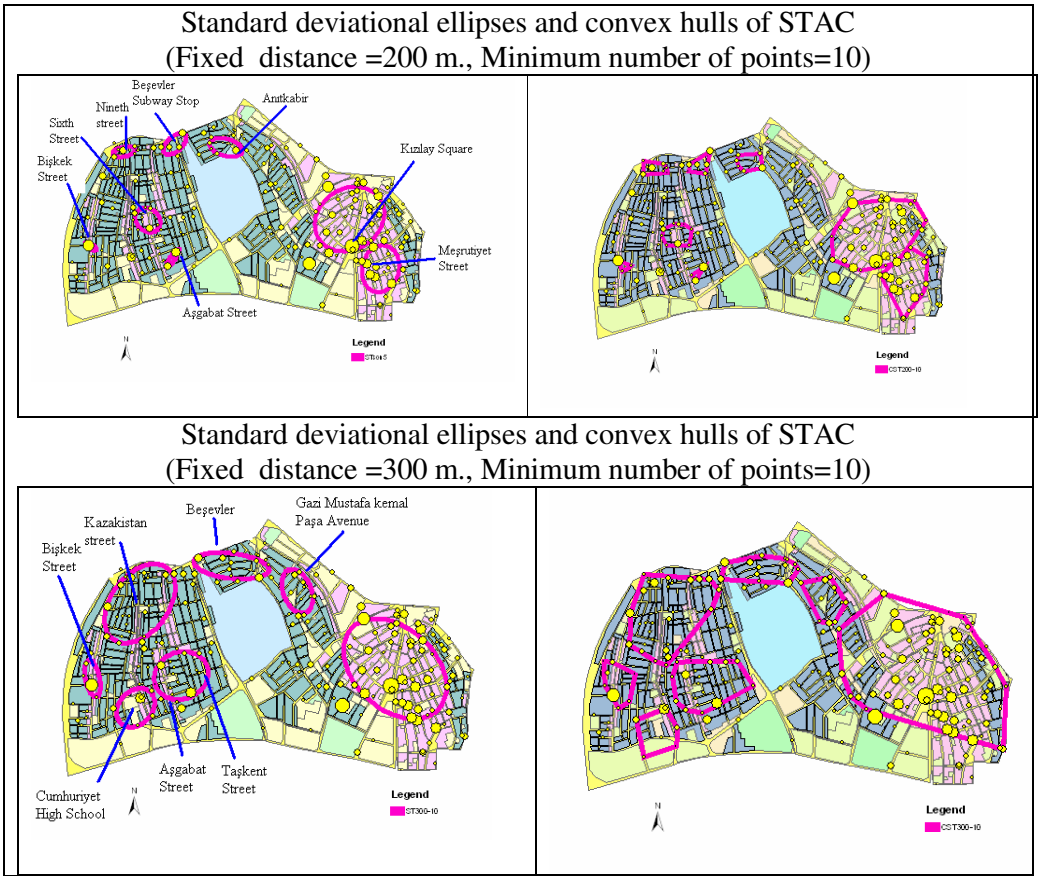
Another interesting point is that small clusters indicate specific areas prone to a specific crime type like Olgunlar and property theft while big clusters include several types of crime in a larger area.

When fixed distance is 300 m., a lot of clusters are generated as expected. Almost all specific areas subject to crime are represented by a cluster. For example, there are clusters at Kızılay square, several parts of Aşgabat Street, Ministry of Health.

When fixed distance increases to 400 m. and 500 m., some clusters are combined and sizes are increased at both side of the study area. The clusters change in orientation and especially at size, when the distance becomes 600 m. Clusters in Bahçelievler, Emek and Beşevler, Maltepe combined and represented by only one cluster. Also, 3rd road in Namık Kemal Street is covered by one cluster.

### 4.3. STAC Hot Spot Areas

STAC is another crime hot spot program which is quick, visual and easy to use (Levine, 2002). STAC identifies the major concentrations of points for a given distribution. Circles are drawn and overlaid for points in a defined grid. Circles having more number of points are ranked and drawn until no overlapping circles exist. After trying lots of combinations, four combinations of fixed distance and "10" minimum number of points mapped above as given in Figure 4.5. Reasons of choosing 4 combinations are when fixed number exceeds 400 m., there is only one cluster and less than 200 m. there are no clusters at Bahçelievler police precinct. Also, when fixed distance is 200 m., the clusters especially in the west side are inconsiderably small. Hence, the map does not give valuable information about the densely populated crime areas. Fixed distance of 400 m. is again gives too big clusters unable to include useful data. Several "minimum number of points" is tried for each fixed distance and 10 are selected to be used in the analysis. Fixed distance of 300 m. is an effective number when compared to the others selected. 5 and 10 "minimum number of points" is tried and no difference is realized. It gives 7 clusters in the study area. Considering the other clustering analysis, 7 is found to be optimal number in this study area with this number of incidents. Also, two visualization techniques are overlaid to point out the difference. As STAC is not restricted to include all the observation, the difference between the standard deviational ellipses and convex hulls is not considerably much as seen in Figure 4.6.



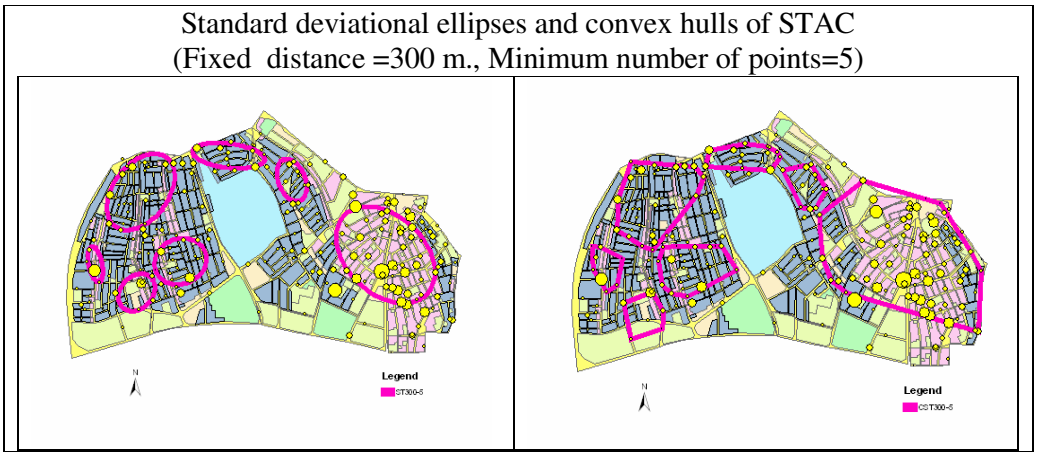




Figure 4.5.STAC hot clusters for the incidents

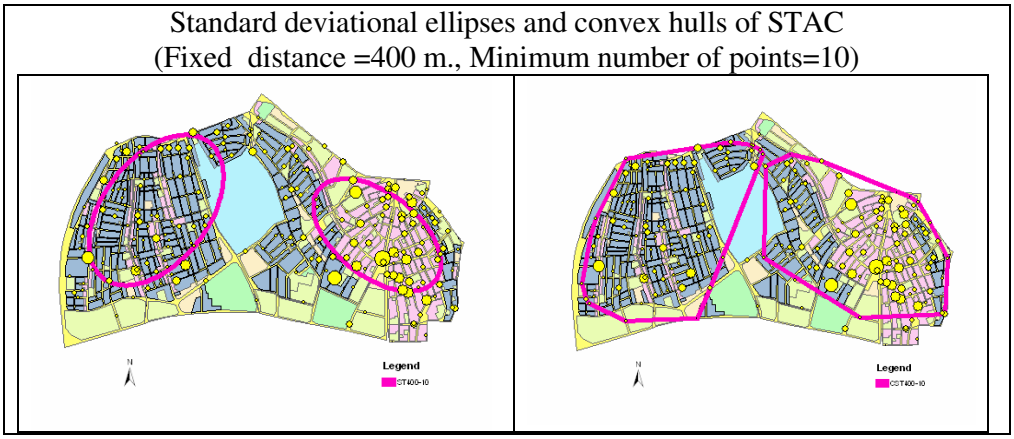



Figure 4.5.STAC hot clusters for the incidents (cont'd)

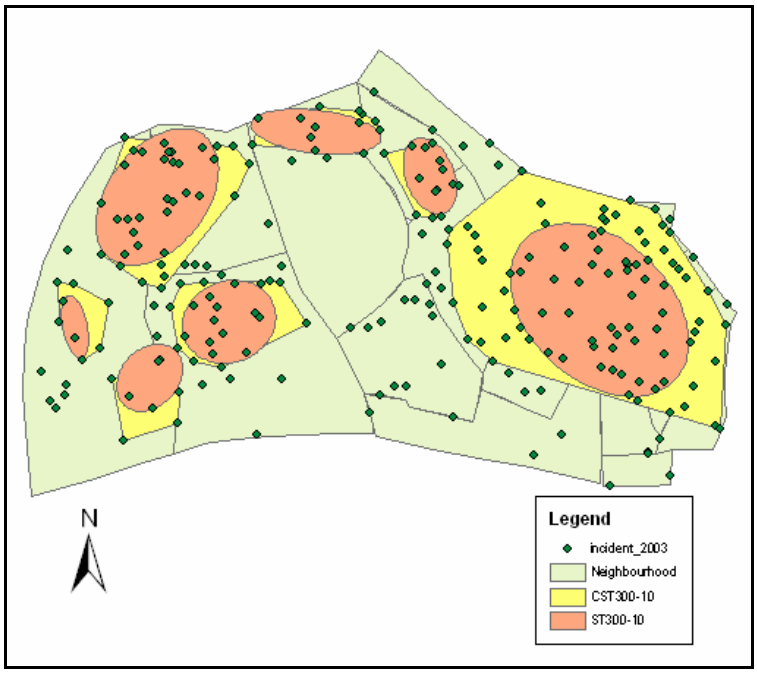



Figure 4.6.STAC representations: standard deviational ellipses and convex hulls for fixed distance 300 m and "minimum number of points" 10.

There are 8 clusters when 200-10 combination is mapped. The first cluster located at the beginning of the Aşgabat Street beside a park area. 2nd cluster covers a part of Bişkek Street and a market area which is generally empty. In the intersection area of 7th and 6th streets, cluster 3 is located. Restaurants, shops and markets are

included by the 3rd cluster. Cluster 4 is lying on the 9th street in Bahçelievler. Clusters in Bahçelievler are located in residential and commercial (fixed) areas. However, cluster 4 is only located in residential areas, mostly prone to burglary indeed. The 5th cluster resides in Beşevler, near to university campuses and the 6th one is beside of the Anıtkabir. Last two clusters are located in Çankaya part in commercial areas. These two clusters have more areas than the rest. One is covering Necatibey, Mithatpaşa Streets and Kızılay Square, whereas the other is covering the area near to Meşrutiyet Street.

All the size of the clusters increases when the fixed distance becomes 300 m. This is an expected result because of the increase in the search area. The first cluster is relatively a big cluster including the area between Aşgabat and Taşkent Streets. In addition, the other part of the Aşgabat Street has another cluster containing Cumhuriyet High School. The third and the fourth clusters are located near Bişkek and Kazakistan Streets. All the four clusters are located in residential areas where there are also commercial areas at two sides of the streets. Fifth cluster covers Beşevler, which includes residential areas. The sixth cluster is interesting as there is more than average number of auto theft in that area, which is between Gazi Mustafa Kemal Paşa Avenue and Turgut Reis Street which is mostly residential. Two clusters in Çankaya in the previous combination become one cluster covering almost all the commercial areas in Çankaya. The last combination 400-10 has two clusters, located in both side of the police precincts.

### 4.4. ISODATA Clustering

ISODATA classification is applied with TNTmips 6.4 software to the study area. The classification is similar to the K Means method but incorporates procedures for splitting, combining, and discarding trial sub regions as it calculates the optimal set of sub regions (Web9). In the software, the desired number of classes, minimum number of cluster cells, maximum standard deviation, and minimum distance (desired diameter) can be selected to group the data in sub regions. However, the weak point of the software is that the results are not always

reflecting the same properties with the options selected. To reach the desired number of groups all the settings should be tried and balanced. It is unable to get results as standard deviational ellipses with TNTmips 6.4. The results are available with convex hulls. Three number of classes (4, 6 ,7) are mapped to ease the comparison with the other methods and get more meaningful representations. The minimum bounding polygons cover all the observation points in the area. This method results in more partitioning in Merkez Çankaya police precinct when the number of classes increases.

To explain the relationship between land-use and clusters, land-use of the study area is mapped (Figure 4.7) and the name of the neighborhoods are illustrated in Figure 4.8. 4 clusters divide the area forming Emek- Yukarı Bahçelievler, Beşevler-Bahçelievler, Anıttepe-Yücetepe-Maltepe and Kızılay-Meşrutiyet-Kocatepe-Fidanlık regions. When the cluster number rises to 6, Bahçelievler, Emek, Yukarı Bahçelievler sub-divided into 3 from 2 clusters. Also, cluster in the mid of the area is divided into 2. The difference between the six and the seven clusters is the division of Sağlık and Korkut Reis neighborhoods. The reason can be Sıhhiye bazaar and intersection of vital roads which can be explained by crime pattern theory.

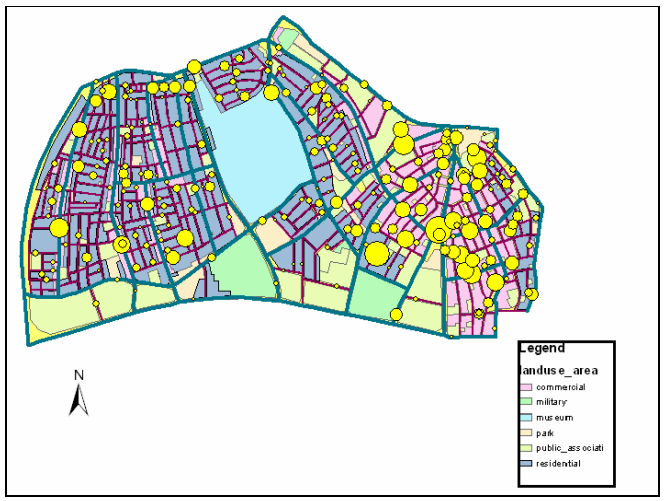



Figure 4.7. Land-use area map of Bahçelievler and Merkez Çankaya police precincts

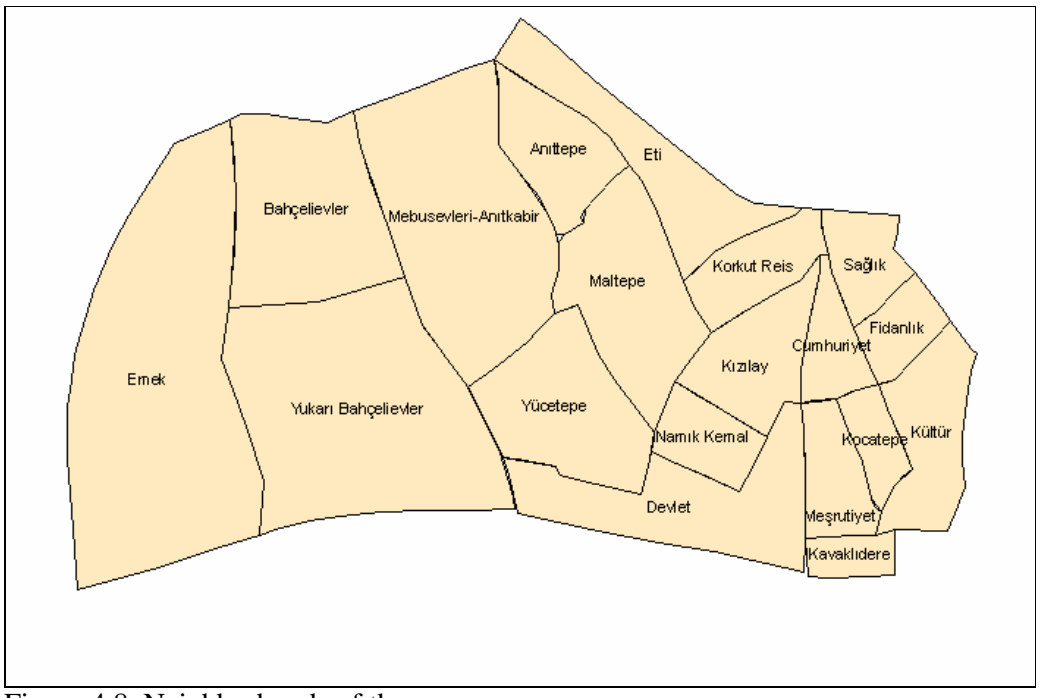


Figure 4.8. Neighborhoods of the area

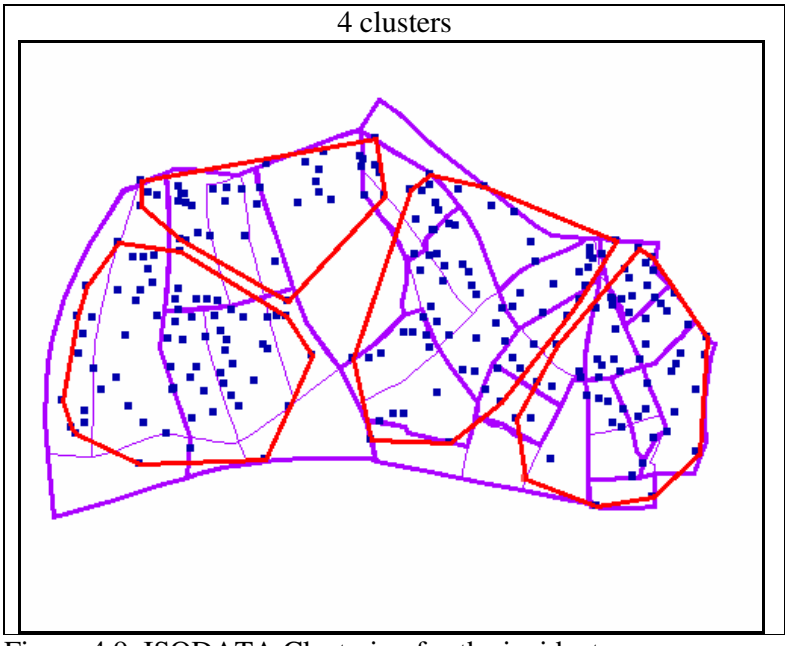


Figure 4.9. ISODATA Clustering for the incidents

Figure 4.9. ISODATA Clustering for the incidents (cont'd)

### 4.5. Fuzzy Clustering

The Fuzzy Clustering method to generate cluster regions uses fuzzy logic concepts to calculate sub regions based on the distance which is the desired average diameter specified by user. Fuzzy clustering is generalized partitioning method differing in the objective function. All the observations have probabilities of having included in a cluster and assigned to clusters with the highest probability. Fuzzy approach is important in clustering as it evaluates all clusters

individually and give more informative results. In this method, TNTmips 6.4 is employed to generate clusters. 4, 6 and 7 clusters are formed to compare with different methods (Figure 4.10).

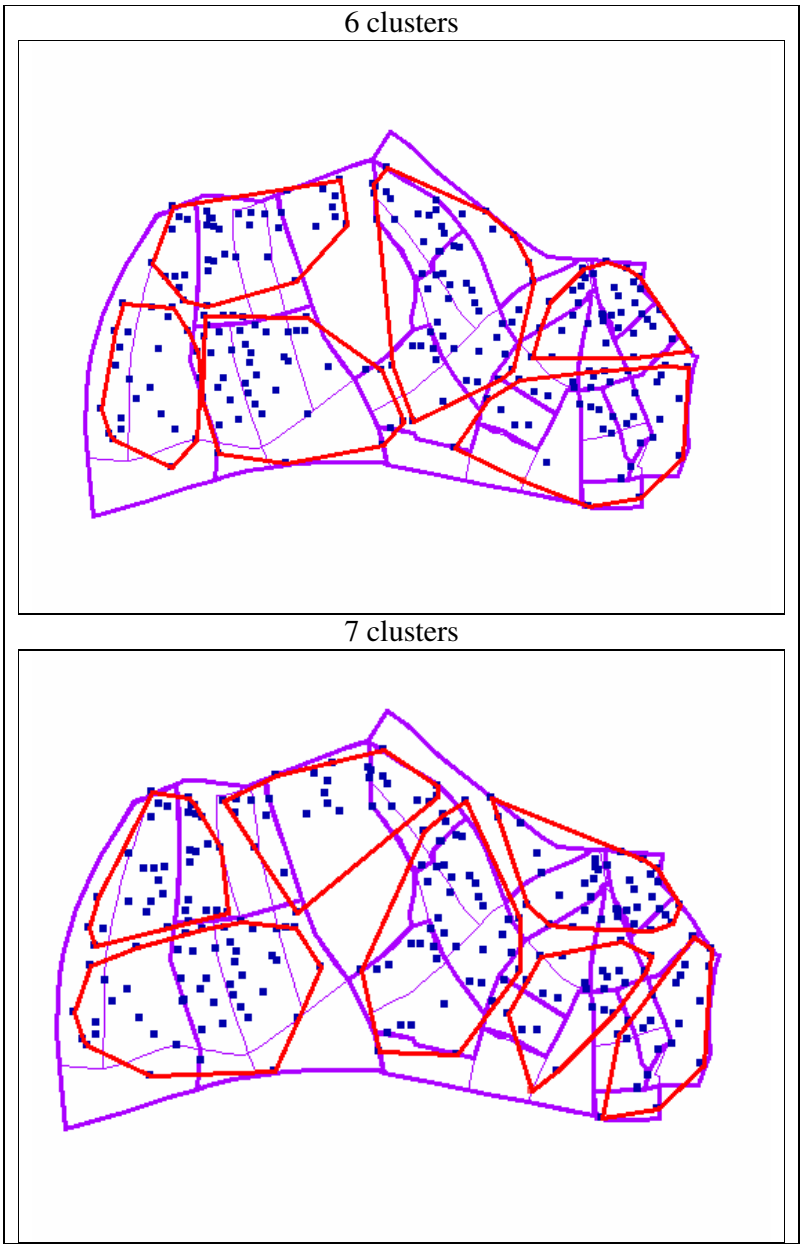



Figure 4.10.Fuzzy clustering for the incidents

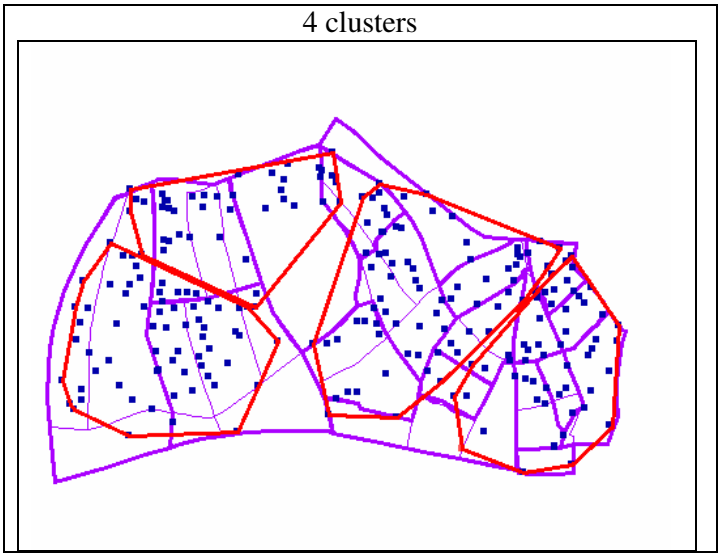


Figure 4.10.Fuzzy clustering for the incidents (cont'd)

The area covered by fuzzy clusters is looking like the area covered by ISODATA clusters and especially 4 cluster configuration is almost the same. When the cluster number increases to 6, the starting point of partitioning is Sağlık neighborhood. At the last configuration Kocatepe-Meşrutiyet and Vatan-Namık Kemal neighborhoods are divided. Neighborhoods are illustrated at Figure 4.8. Those are the places where the crime rates are high rather than the Bahçelievler police precinct part of the study area.

### 4.6. Geographical Analysis Machine Approach

Gam/K software is used to get the resulting maps of the analysis (Map 4.11). Results are represented by kernel smoothing approach by the software. As seen from maps, although maximum and minimum search radiuses are changing, the influence area is not affected. The difference of this method to define hot spot areas is consideration of the weight procedure. In this approach population and number of incidents for each neighborhood are recorded and the results are determined with respect to these values. Results indicate that 8 neighborhoods in Merkez Çankaya precinct are significant according to the approach. These neighborhoods are Eti, Korkut Reis, Sağlık, Kızılay, Cumhuriyet, Fidanlık,

Kocatepe and Meşrutiyet. Figure 4.12 demonstrates the land-use and the significant areas. All the neighborhoods have commercial areas except Fidanlık and Sağlık neighborhoods which have both residential and commercial areas. The results are not unforeseeable that commercial areas have lower population than residential areas in the area. The important question arises here that if the smoothed areas are really hot spots. The answer can be, some of them are and some of them are not. However, to decide about a hot spot, area should be known and investigated carefully.

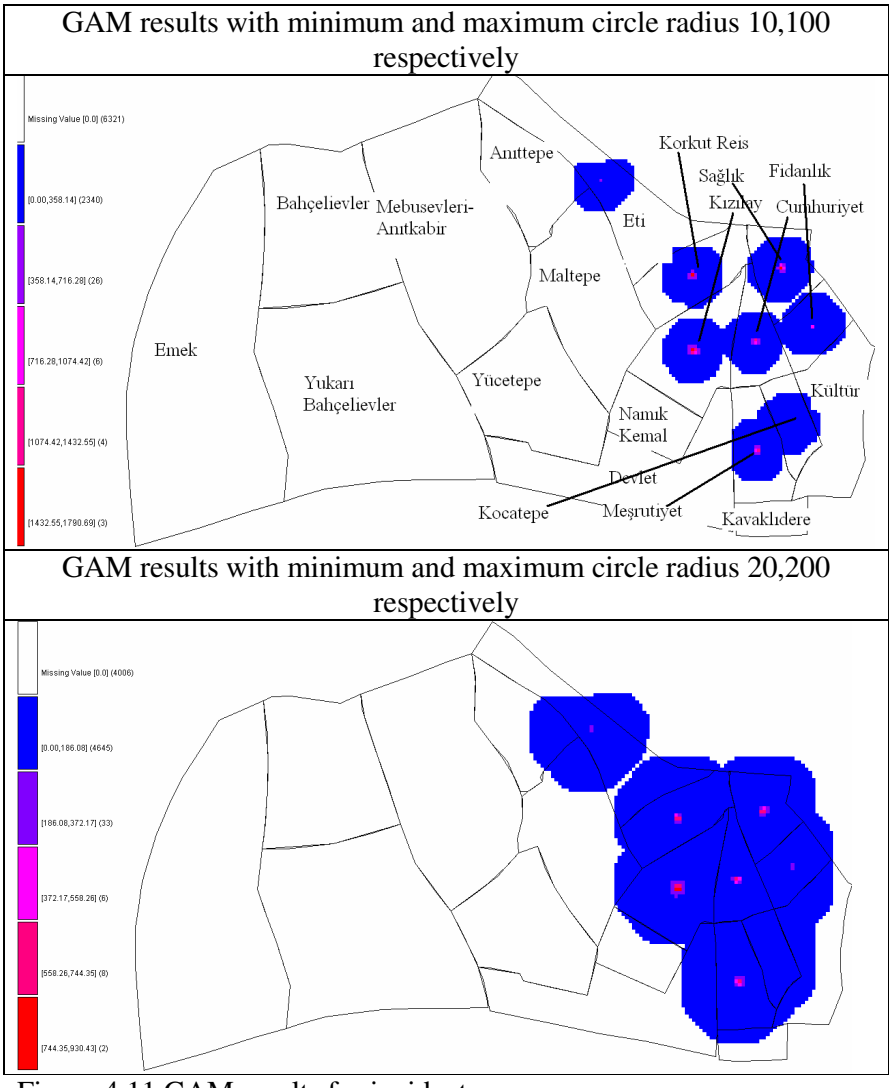


Figure 4.11.GAM results for incidents

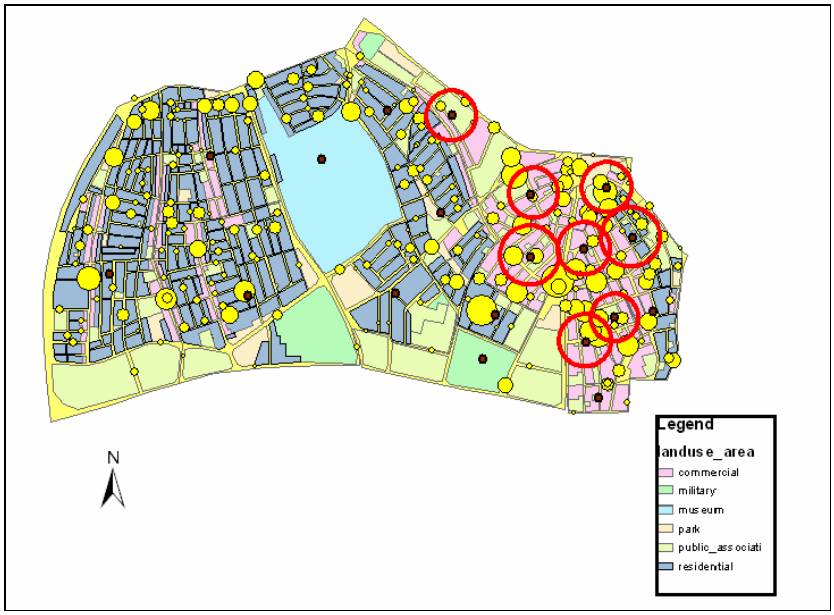


Figure 4.12.Gam clusters with land-use area.

### 4.7. Comparison of the clustering methods

The aim to compare the clustering methods stated in this chapter is to choose the most appropriate for the spatio-temporal crime prediction model. There are several reasons to choose a clustering algorithm but the most important criteria is to select the algorithm according to the purpose of usage. Firstly, to make a general comparison between the clustering methods in the study area, convex hull (7 clusters) maps are represented (Figure 4.13). Briefly, K-means, ISODATA, and fuzzy clustering methods are types of partitioning approach and cover all the observations in the area. Nnh hierarchical clustering is distance specific hierarchical approach and STAC is combination of two approaches, partitioning approach with search circles and hierarchical approach with aggregating smaller clusters into larger clusters.

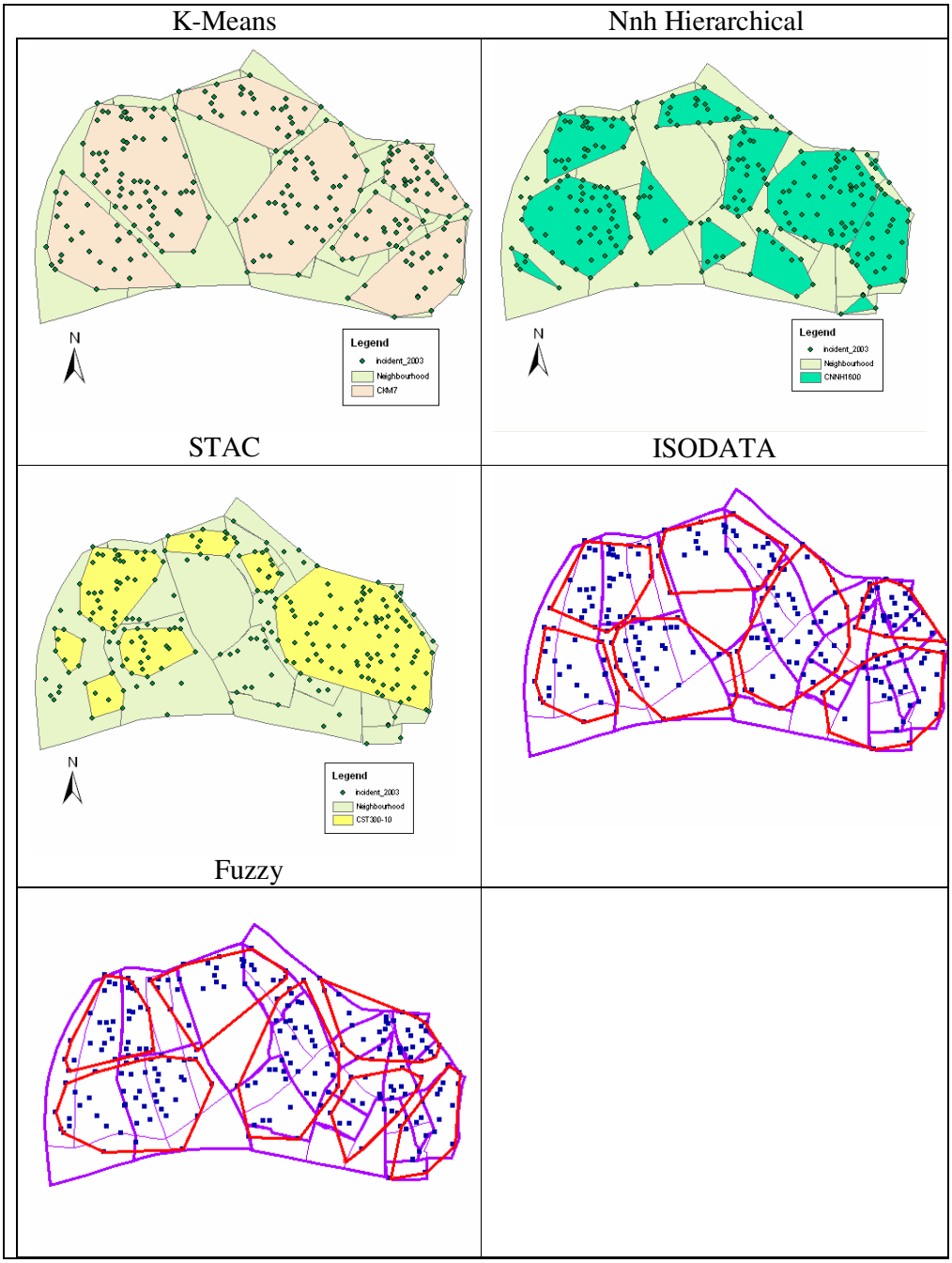



Figure 4.13.Resulting maps of the clustering methods

To compare the partitioning methods, K-means and fuzzy clustering methods are similar in Merkez Çankaya precinct in general as seen in Figure 4.13, evidence of better representation of observations at that part of the study area. Both methods work with optimization procedure, where fuzzy clustering concerns the

possibilities, whereas K-means has hard partitioning. As fuzzy gives possibilities to each observation and assign them to clusters with highest possibility, the sub-division of the incidents at that part are more accurate. ISODATA is one type of optimization based clustering but has different orientation than the other two (Figure 4.13). Especially in Bahçelievler region, ISODATA has more partitioning. ISODATA is indeed a classification method used for image processing and is not commonly used in criminological issues as stated in methodology. Therefore it is found in this study that ISODATA clustering is not appropriate.

In partitioning based clustering methods, the number of clusters is defined by the user. This can be an advantage or a disadvantage according to the purpose of usage. Inclusion of all the points is one of the limitations of this approach. Spatial outliers are forced to be included to clusters, hence cluster orientation and sizes are deviating from the optimal. Implementation of partitioning approach is difficult than the other approaches because it includes an optimization procedure. Also, K-means objective function is not linear as the distance metric is squared Euclidean so heuristic approaches are used to solve the problem. However, besides being difficult to implement, the approach is commercially available and common.

Partitioning based clustering algorithms are preferably used to allocate resources effectively. For example if there are 4 main teams available, dividing the area to 4 parts will be appropriate for an effective usage. Also, to look at the general view of the area, this approach could be applicable.

The first order clusters of Nnh hierarchical approach is too many to evaluate the general perspective. The method is useful when the area concerned is relatively small like a street segment or a specific place. To detect the density of the crime activities in that area, Nnh hierarchical approach is selected. One of the disadvantage to use the method is recovery which is impossible in following phases when an error occurred in one phase.

STAC and Nnh hierarchical routines are similar compared with the other methods. As STAC includes some form of hierarchical approach, this is not an unexpected result. Both methods divide Bahçelievler region more than the other algorithms. Looking at the statistical results with one nearest neighbor distance in Table 4.2, it is observed that the mean distance in Merkez Çankaya is smaller than Bahçelievler. As Nnh and STAC methods consider the nearest distances, the partitioning of the clusters in Bahçelievler police precinct is meaninful. Also, Merkez Çankaya region has lower test statistics (Z) indicating more clustering in the eastern part.

Table 4.2.Statistical results of nearest neighbor distances of two police precincts.

| | Merkez Çankaya | Bahçelievler |
|---|---|---|
| Mean Nearest Neighbor Distance | 1.45 m | 2.76 m |
| Standard Dev of Nearest Neighbor Distance | 13.76 m | 22.03 m |
| Minimum Distance | 0 m. | 0 m. |
| Maximum Distance | 2892.72 m. | 2534.98 m |
| Nearest Neighbor Index | 0.0449 | 0.0688 |
| Standard Error | 0.47 m. | 0.85 m |
| Test Statistic (Z) | -65.2209 m. | -44.0350 m |
| p-value (one tail) | 0.0001 | 0.0001 |
| p-value (two tail) | 0.0001 | 0.0001 |

STAC and partitioning approaches have quiet different orientations especially in the west and east side of the regions, which have totally dissimilar configurations. STAC have inclined to partition in Bahçelievler, whereas K-means clusters are distributed in Çankaya. The difference is meaningful as STAC algorithm considers the distance measures within observations and mean nearest neighbor distance between observations (Table 4.2) is smaller in Çankaya. To get the optimal result K-means clustering approach should divide the data into groups in Merkez Çankaya police precinct. The reason under this is the minimization procedure of the distance between center and the observations.

According to these discussions STAC is selected to be used in the crime prediction model for several reasons:

1. Clusters of STAC do include more homogenous areas than the other methods. The biggest cluster in Çankaya, almost cover all the commercial area in Çankaya region. Homogeneity of land use in the clusters is an advantage as the crime incidents happened is more typical. Actually, it is an advantage in crime prediction as when number and place of crime incidents are forecasted also crime types will be predicted. Police for example, use the advantage to control the similar areas. STAC is not restricted to include all the observations hence STAC is able to indicate denser crime areas than other methods. This is important in crime prevention for allocating resources effectively. If all the area is going to be searched, there is no meaning to form crime prediction models.

2. The second advantage of STAC is computation efficiency, which is faster than the other methods. Fuzzy and ISODATA can not be represented by standard deviational ellipses and is out of scope at the beginning of the comparison. SDE are preferred to convex hulls not to search all of the area. Also, as opposed to K-means, STAC clusters consist of most of the land marks like schools, shopping centers, sports fields in the area although is not covering all the observations.

Three combinations of STAC are tried. Crime prediction model is daily based model. Hence, data is divided into seven weekdays and clusters are generated for each day. In crime prediction model, areas of clusters should not be so small or so large to get more meaningful results. Areas of clusters in Table 4.3 show that the areas of 200-10 combination is so small like 6789 $m^2$ to choose for control area and 400-10 is so large. Also 300-10 covers % 87 of the observations, while 200-10 covers % 65. Of course, 400-10 has almost all the observations, but cluster sizes are insignificantly big. As a result STAC 300-10 combination is selected for crime prediction model to get more informative and accurate results.

Table 4.3. STAC cluster's density values

| | 200-10 | | | 300-10 | | | 400-10 | | |
|---|---|---|---|---|---|---|---|---|---|
| Clusters | Area($m^2$) | Points | Density | Area($m^2$) | Points | Density | Area($m^2$) | Points | Density |
| 1 | 478455 | 634 | 0.0013 | 812339 | 1031 | 0.0013 | 1075248 | 1179 | 0.0011 |
| 2 | 188878 | 277 | 0.0015 | 432042 | 201 | 0.00047 | 1575484 | 636 | 0.0004 |
| 3 | 66213 | 61 | 0.0009 | 218770 | 119 | 0.00054 | | | |
| 4 | 43410 | 59 | 0.0014 | 159273 | 99 | 0.00062 | | | |
| 5 | 23388 | 55 | 0.0024 | 108534 | 77 | 0.00071 | | | |
| 6 | 61969 | 53 | 0.0009 | 43979 | 64 | 0.0015 | | | |
| 7 | 11446 | 50 | 0.0043 | 120867 | 55 | 0.00046 | | | |
| 8 | 6789 | 44 | 0.0065 | | | | | | |

# CHAPTER 5

## EFFECTS OF DISTANCE METRICS TO STAC ALGORITHM

Clustering algorithms are applied with different distance metrics if the type of distance is not specifically determined in the algorithm. Three types of distance metrics are considered in clustering algorithms: Euclidean, which calculates the panoramic distance, is the most commonly used and preferred distance in clustering analysis. The other type that calculates rectilinear distance is Manhattan and also, squared Euclidean distance is used in clustering algorithms. The aim of this chapter is to detect the influences of different distance metrics to previously selected STAC algorithm. STAC with fixed distance 300 m. and "minimum number of points" 10 is chosen to use in spatio temporal crime prediction model but the distance metric is applied as Euclidean to compare with the other clustering methods in the previous chapter. Since crime prediction model is going to be generated for both Euclidean and Manhattan distance metrics, it will be useful to detect the effect of the structure of two distance metrics to STAC 300-10 combination.

Several distance metrics are in the scope of clustering issue. However, all the algorithms do not let to apply different distance metrics. In fact, K-means clustering minimizes the squared Euclidean distance between the center and the observations and it is not possible to change the metric not to destroy the structure of the algorithm. Software limitations do not allow applying different distance metrics to ISODATA and fuzzy clustering algorithms. STAC and Nnh hierarchical clustering algorithms can be generated with both distance metrics by CrimeStat 3.1, however, as STAC is used in spatio-temporal crime prediction model STAC with fixed distance 300 and minimum number of points 10 is mapped for both of the distance types, Euclidean and Manhattan (Figure 5.1). Also, both methods are mapped in Figure 5.2.

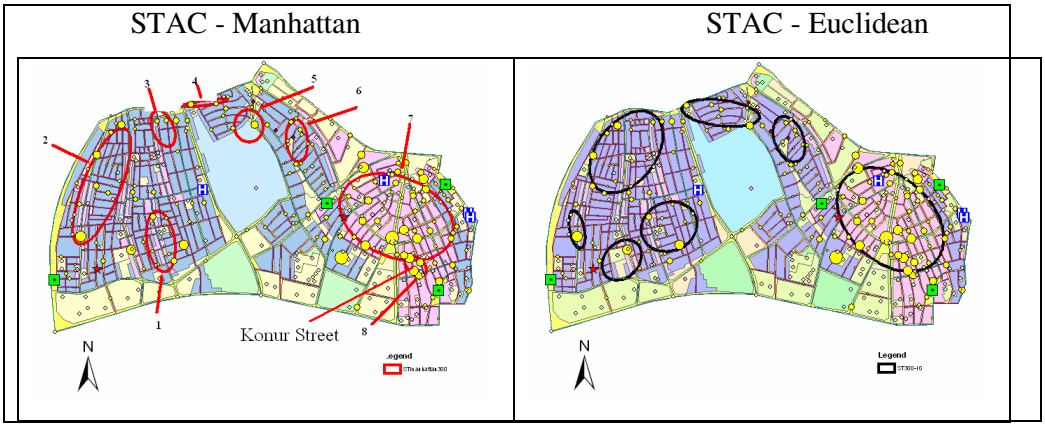
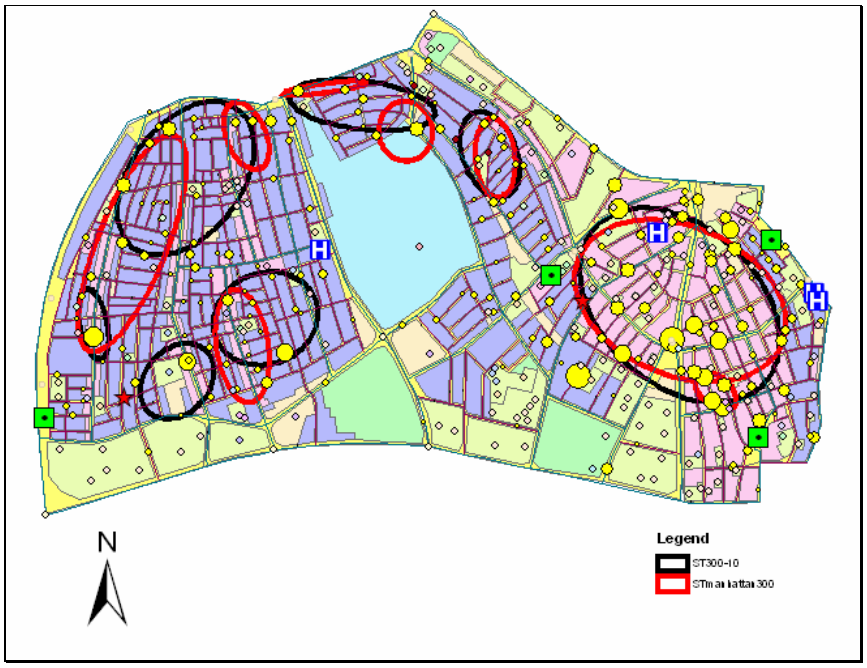



Figure 5.1.STAC clustering with difference distance metric applications.




Map 5.2.STAC clustering with difference distance metric applications at the same map.

Although the number of clusters does not change significantly in STAC hot clusters, the size and the orientation change especially in Bahçelievler and Beşevler parts of the study area. The dense crime area in Çankaya region is represented by only one cluster and the different distance applications do not affect the configuration significantly. Only the area near Konur Street is divided

into another cluster which is quite small. However; in Bahçelievler, the sizes and the shapes of the clusters updated in order to include points which the Euclidean and the Manhattan distances of the observations are similar in Manhattan distance application. First cluster of STAC with Manhattan distance metric shrinks and become elongated to north-south direction that is depicted in Figure 5.1. It covers all the Aşgabat Street. However, the cluster generated by STAC with Euclidean distance metric in the same location is spread on the residential areas. Also, cluster on Cumhuriyet high school is not represented by STAC with Manhattan distance metric. The upper part of Emek and Yukarı Bahçelievler are assigned to only one cluster in STAC with Euclidean distance metric, however, cluster of STAC with Manhattan distance metric includes the area between Bişkek and Kazakistan Streets lengthening to north-south direction. Cluster in Beşevler is separated into two clusters: one is located toward the university area and the other is located in the intersection area of Anıt and Gençlik Streets. At last, cluster in Anıttepe gets smaller with STAC with Manhattan distance metric. As a result, STAC with Manhattan distance metric is mostly located in the area elongated to the streets in order to decrease the Manhattan distance of the points in the clusters.

Examining more structural part such as area, number of incidents and road density, several results can be obtained. Hence, total area (Table 5.1, 5.2), and road densities are calculated (Table 5.3, 5.4). The area covered by clusters is larger in Euclidean distance (1.89 $km^2$) is greater than Manhattan clusters (1.49 $km^2$) (Table 5.1, 5.2). Also the number of incidents represented is higher in Euclidean than Manhattan distance. The proportion of number of incidents in clusters to total number of incidents makes STAC with Euclidean distance metric clusters are more representative than STAC with Manhattan distance metric except usurp.

Table 5.1.Total area of clusters for STAC clustering.

| Clustering methods | Total area of clusters |
|---|---|
| STAC with Euclidean distance metric | 1.89 $km^2$ |
| STAC with Manhattan distance metric | 1.49 $km^2$ |

To discuss the crime clusters in terms of distance metrics, point densities are also important to decide which is going to be used. If the point density is too low, it is not informative to have the dense crime areas. Lower values under 0.0005 are highlighted in Table 5.2 and Table 5.3 and there is little difference between the values of two distance metrics. However, when the density values are summed, Manhattan is calculated two times dense than Euclidean. Also, values of road densities confirm the algorithms as Manhattan has lower road density as it elongated in the area in east-west direction.

Table 5.2. Point and road densities of STAC Manhattan clusters

| Cluters | Area (m2) | Points | Density | Road lenght(m) | Road density |
|---|---|---|---|---|---|
| 1 | 132769 | 74 | 0.00056 | 2274 | 0,017 |
| 2 | 362710 | 153 | **0.0004** | 6946 | 0,019 |
| 3 | 63191 | 31 | **0.00049** | 1008 | 0,016 |
| 4 | 17352 | 35 | 0.0020 | 127 | **0,007** |
| 5 | 72699 | 33 | **0.00045** | 750 | **0,010** |
| 6 | 70554 | 58 | 0.00082 | 1202 | 0,017 |
| 7 | 737335 | 761 | 0.0010 | 12396 | 0,017 |
| 8 | 30702 | 143 | 0.0046 | 289 | **0,009** |
| Total | | | **0.010** | | **0.112** |

Table 5.3. Point and road densities of STAC Euclidean clusters

| Cluters | Area (m2) | Points | Density | Road lenght(m) | Road density |
|---|---|---|---|---|---|
| 1 | 218770 | 119 | 0.00054 | 4041 | 0,018 |
| 2 | 120867 | 55 | **0.00046** | 1479 | 0,012 |
| 3 | 43979 | 64 | 0.0015 | 625 | 0,014 |
| 4 | 432042 | 201 | **0.00047** | 7966 | 0,018 |
| 5 | 159273 | 99 | 0.00062 | 2001 | 0,013 |
| 6 | 108534 | 77 | 0.00071 | 1624 | 0,015 |
| 7 | 812339 | 1031 | 0.0013 | 13013 | 0,016 |
| Total | | | **0.0056** | | **0.106** |

To sum up, two different clustering metrics are used in spatio temporal crime prediction model. Therefore, structural differences of two distance metric

applications of STAC is analyzed and evaluated. The area that both types of clusters represent is sometimes similar and sometimes different with respect to the land-use. One of the limitations here is not to be able to make the analysis with original network because of the software. In fact, the values of Manhattan distance reflect the original network more than the Euclidean distance. The reason behind this is that the road networks in reality are formed by rectilinear distances. Similarity of the original network does not mean to be more representative in crime prediction model. To understand which algorithm better fits, spatial dissagregation method should be applied, error terms should be calculated and the clusters and predictions for each cluster are evaluated.

# CHAPTER 6

## SPATIO-TEMPORAL CRIME PREDICTION MODEL WITH ARIMA MODEL FITTING AND THE SIMPLE SPATIAL DISAGGREGATION APPROACH

Forecasting is gaining popularity in crime with the advances in technology. Predicting the number of crimes, the influence area, the time, and the type of crime enable to overcome the occurrence of crime. Several ways can be concluded in crime forecasting such as hot spots, time series analysis and various statistical models. In this chapter, a spatio-temporal crime prediction model is generated with ARIMA forecasting and spatial disaggregation approach. A Box-Jenkins ARIMA model is commonly used in several sciences like in economics, biology, production planning, etc. The ARIMA model has four step iteration; identification, estimation, diagnostic checking and forecasting. Forecasted values are disaggregated into the area by spatial disaggregation approach. To implement spatial disaggregation approach, area should be divided into meaningful parts. For this reason, STAC clustering model is selected and predicted values are assigned to these week-day clusters for Euclidean and Manhattan distance metrics. The aim is to form a spatio-temporal crime prediction model and test which distance metric is better fit to the model.

To predict the future values, Box-Jenkins ARIMA model is fit to daily data for the year 2003. All the steps are evaluated iteratively and forecasted values are gained. Minitab, dataplot and Xlstat are employed during these processes and Microsoft Excel is used in statistical calculations. The following part of this study is applying spatial disaggregation approach to the clusters investigated in the previous chapter.

As daily forecasts are found, spatial disaggregation approach is applied to days of the week. STAC is selected as the clustering model and seven days of the week

are clustered for two distance metrics; Euclidean and Manhattan. To understand how the model fit the data, spatio-temporal root mean square estimate is calculated for the entire model. At last, forecasted values are disaggregated into the daily STAC with Euclidean distance metric and STAC with Manhattan distance metric clusters.

### 6.1. Fitting Box-Jenkins ARIMA model to daily number of incidents data

The first stage in the Box-Jenkins model is the identification stage. In order to tentatively identify a model, first whether the time series is stationary or not should be determined. A time series is stationary if the statistical properties like mean and variance are essentially constant over time (Boverman and O'Connell, 1993). The simplest way to understand this is to plot the values against time. If the values seem to fluctuate with constant variation around a constant mean, it is reasonable to believe that the time series is stationary. Plotting the number of incidents of each day against time, time series plot of number of incidents in Figure 6.1 is gained. Although having some outliers especially in the second half of the year, the graph seems stationary. There is no evidence of a trend and seasonality in the data. As days of the weeks are used, week periods are more prone to indicate seasonality.

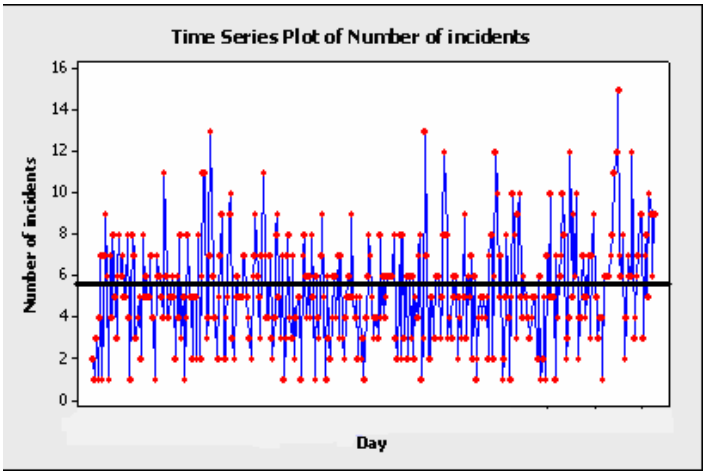


Figure 6.1.Time series plot of number of incidents

In order to utilize more sophisticated results, there are several ways to evaluate the stationarity of the time series. One of the ways is to apply unit root tests. Phillips-Perron test is applied to the data to confirm stationarity of the time series data. The test results are:

H0: Unit root; H1: Stationarity
Alpha = 0.1705
Test statistic: -322.67
p-value = 0.00000
5% Critical region: < -14.51
10% Critical region: < -11.65

When a < 1 the process is stationary (Phillips and Perron, 1988). According to the test result H0 is rejected in favor of H1, at the 5% significance level, which that means the process is stationary.

In stationarity, there are two concepts which should be considered; mean and variance. Non-stationarity can be transformed to stationarity with respect to these concepts. If the problem is caused from mean, differencing should be applied or if is caused by variance, transformation should be done. To detect the mean and variance movements, both of them are plotted with dividing the data into 8 lags. Movement of mean in 8 lags is not so volatile; the values are between 4, 75 and 5, 25 until the $8^{th}$ lag (Figure 6.2). Hence, the variation of the mean is not significant. Also, looking at the autocorrelation plot, stationarity can be evaluated. Looking at Figure 6.3, as the values are reaching 0, time series of number of incidents do not need differencing (Web7). Also, to search the seasonality autocorrelation function is very important. The values over the red line means, autocorrelation function is significant at that lag. Lag 1 and lag 4 indicates significance according to the Figure 6.3, however, there is no evidence of seasonality in the data.

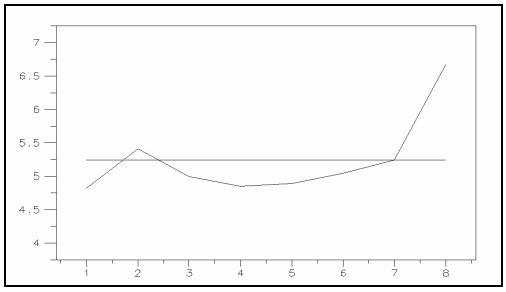


Figure 6.2. Variation of the mean plot of time series data

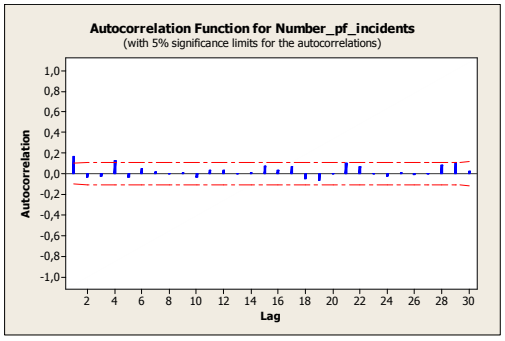


Figure 6.3.Autocorrelation plot of number of incidents

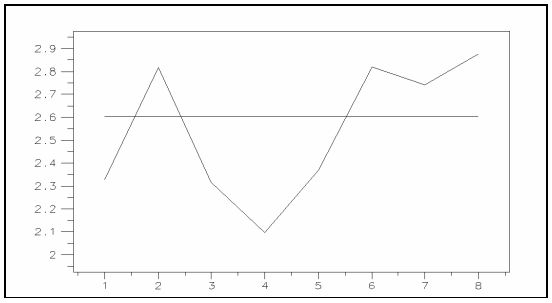


Figure 6.4.Variation of variance plot of time series data

To seek the movement, variation of variance is plotted for the time series data. Mean is 2.6 and maximum deviation of the mean is 0.5, which does not seem significant.

Histogram of the data indicates that the data seem normally distributed (Figure 6.5).

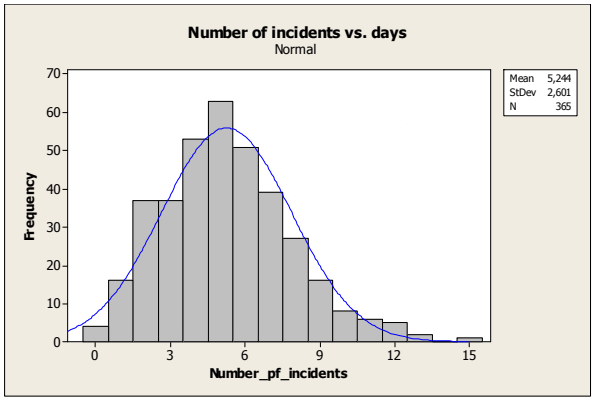


Figure 6.5. Histogram of the data.

After, confirming the stationarity of the data, the next step is to determine the levels of AR and MA values. As there is no need to difference the data, the model is turned to an ARMA model as I represent the amount of differencing. For the decision of the levels, autocorrelogram and partial autocorrelogram are plotted and as it is an iterative process, different combinations are tried to get the best result.

Autocorrelogram and autocorrelation function values are not only important to detect stationarity but also good indicators to determine the level of MA(moving average) level (Web8). As seen obviously from the Figure 6.6 that the lags are significant when the lag values pass the red line. Red line indicates the 5% significance level of autocorrelations. Another and more informative evidence is to look at the autocorrelation values and t statistics. Bowerman and O'Connell (1993) stated that for lower lags (lag < 3); the spike exists if t value is greater than 1.6 and for higher lags, a spike is considered to exist if t is greater than 2. According to this statement, it is convenient to say according to Table 6.1 that lag 1 and lag 4 are significant.

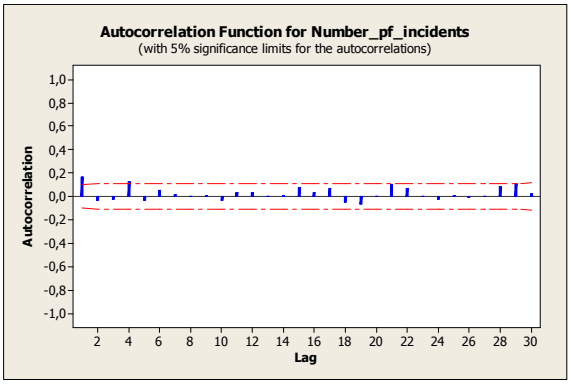


Figure 6.6. Autocorrelation plot of number of incidents

Table 6.1.Autocorrelation function and t values of each lag

| Lag | Autocorrelation function | t value |
|---|---|---|
| 1 | 0,169481282 | **3,237935** |
| 2 | -0,030176045 | -0,56063 |
| 3 | -0,023595829 | -0,438 |
| 4 | 0,123941384 | **2,29949** |
| 5 | -0,029842495 | -0,54582 |
| 6 | 0,047166726 | 0,861975 |
| 7 | 0,015566709 | 0,283905 |
| 8 | -0,004141469 | -0,07552 |
| 9 | 0,005577428 | 0,101697 |
| 10 | -0,03546866 | -0,64671 |
| 11 | 0,03645105 | 0,663857 |
| 12 | 0,0346051 | 0,629479 |
| 13 | 0,00077731 | 0,014124 |
| 14 | 0,007456513 | 0,135489 |
| 15 | 0,072078958 | 1,309653 |
| 16 | 0,032557722 | 0,588804 |
| 17 | 0,066941173 | 1,209477 |
| 18 | -0,053455023 | -0,96197 |
| 19 | -0,063539175 | -1,14055 |
| 20 | -0,003481274 | -0,06227 |
| 21 | 0,09747957 | 1,743569 |
| 22 | 0,070147135 | 1,244366 |
| 23 | 0,000396916 | 0,007011 |
| 24 | -0,022271987 | -0,39343 |
| 25 | 0,01107154 | 0,195491 |
| 26 | -0,004894023 | -0,08641 |
| 27 | 0,004091922 | 0,072242 |
| 28 | 0,086051627 | 1,51921 |
| 29 | 0,108454261 | 1,902727 |
| 30 | 0,028862021 | 0,501408 |

In partial autocorrelation values, the same principle is valid but for all lags the t value should be greater than 2 to consider a spike (Bowerman and O'Connell, 1993). Both the partial autocorrelogram and the Table 6.2 point out that again lag 1 and lag 4 are significant. Partial autocorrelation function is important to decide the level of AR (Autoregressive) part in the process.

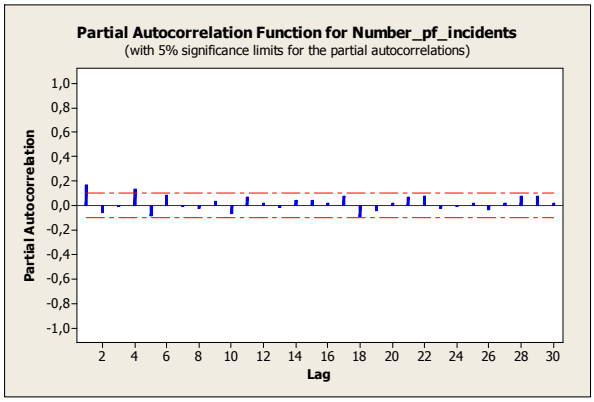


Figure 6.7.Partial autocorrelation function plot of number of incidents

Table 6.2. Partial autocorrelation function and t values of each lag.

| Lag | Partial autocorrelation function | t value |
|---|---|---|
| 1 | 0,169481282 | **3,237935** |
| 2 | -0,06064182 | -1,15856 |
| 3 | -0,008157204 | -0,15584 |
| 4 | 0,132040942 | **2,522639** |
| 5 | -0,080718923 | -1,54213 |
| 6 | 0,081312993 | 1,553483 |
| 7 | -0,005055487 | -0,09658 |
| 8 | -0,024180242 | -0,46196 |
| 9 | 0,033806948 | 0,645881 |
| 10 | -0,067822365 | -1,29574 |
| 11 | 0,065799919 | 1,257106 |
| 12 | 0,015038006 | 0,287301 |
| 13 | -0,018965355 | -0,36233 |
| 14 | 0,039498259 | 0,754613 |
| 15 | 0,044783468 | 0,855587 |
| 16 | 0,015871 | 0,303215 |
| 17 | 0,072922356 | 1,39318 |
| 18 | -0,094567672 | -1,80671 |
| 19 | -0,040220599 | -0,76841 |
| 20 | 0,01583234 | 0,302476 |
| 21 | 0,065147214 | 1,244636 |
| 22 | 0,071214916 | 1,360559 |
| 23 | -0,026175376 | -0,50008 |
| 24 | -0,009572422 | -0,18288 |
| 25 | 0,017506305 | 0,334457 |
| 26 | -0,035229543 | -0,67306 |
| 27 | 0,020374169 | 0,389248 |
| 28 | 0,07353125 | 1,404813 |
| 29 | 0,073041585 | 1,395458 |
| 30 | 0,020679518 | 0,395082 |

After detecting spikes existing in graphs, which will guide in trial period, several combinations of AR and MA levels is going to be evaluated to get the best result. As spikes are detected at lag 4 for autocorreloram and partial autocorreloram, the trial starts from AR(4) and MA(4).

Probability value is an indicator to detect the statistical significance of the model. If the probability value is smaller than 0.05, the parameter is significant in 95% significance level. When the parameter is significant it should be involved into the model. Standard squared error of residuals is the second issue to be considered in

significance of the model. Lower the SS value, higher the accuracy of the model. At last in diagnostic checking part, modified Box-Pierce (Ljung-Box) Chi-Square statistic is going to be evaluated to analyze the residuals obtained from the model. If the probability value is near to the value 1, it is reasonable to say that the model is adequate (Bowerman, O'Connell, 1993). Also, the adequacy of the model should be supported with normal probability plot and the autocorrelogram and partial autocorrelogram of the residuals.

To detect the levels of the model firstly single values of AR and MA levels are calculated and the results are evaluated. In Tables 6.3, 6.5, 6.7, and 6.9 coefficients, t, and p values of parameters are indicated. For all single AR and MA values there is no situation where all the p values are smaller than 0.05. In all the levels residuals sum of square values have slight differences not indicating an improvement between the trials. In Tables 6.4, 6.6, 6.8 and 6.10 Modified Box-Pierce test statistics are demonstrated. p values indicate the accuracy of the model if values are near to 1. As Bowerman and O'Connell (1993) noticed, higher the probability values of Box-Pierce statistics, higher the evidence of adequacy of the model. However, the p values of single AR and MA values are not adequate to prove the model's adequacy.

Table 6.3.Final estimates of parameters AR(4)

| Type | Coefficient | t | p |
|---|---|---|---|
| AR 1 | 0,1812 | 3,47 | 0,001 |
| AR 2 | -0,0508 | -0,96 | **0,340** |
| AR 3 | -0.0303 | -0,57 | **0,570** |
| AR 4 | 0,1352 | 2,58 | 0,010 |
| Constant | 4,0069 | 30,11 | 0 |
| SS | **2340,11** | | |
| MS | **6,48** | | |

Table 6.4.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(4)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 7,7 | 18,5 | 29,4 | 37,9 |
| P-value | 0,364 | 0,487 | 0,549 | 0,692 |

Table 6.5.Final estimates of parameters AR(2)

| Type | Coefficient | t | p |
|---|---|---|---|
| AR 1 | 0,1803 | 3,44 | 0,001 |
| AR 2 | -0,0608 | -1,16 | **0,248** |
| Constant | 4,6162 | 34,47 | 0 |
| SS | **2383,32** | | |
| MS | 6,57 | | |

Table 6.6.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(2)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 13 | 23,7 | 34,7 | 44,9 |
| P-value | 0,162 | 0,306 | 0,386 | 0,477 |

Table 6.7.Final estimates of parameters MA(4)

| Type | Coefficient | t | p |
|---|---|---|---|
| MA 1 | -0,2126 | -4,10 | 0,000 |
| MA 2 | 0,0426 | 0,81 | **0,421** |
| MA 3 | 0,0592 | 1,11 | **0,266** |
| MA 4 | -0,1749 | -3,36 | 0,001 |
| Constant | 5,2411 | 30,83 | 0 |
| SS | **2310,11** | | |
| MS | 6,40 | | |

Table 6.8.Modified Box-Pierce (Ljung-Box) Chi-Square statistic MA(4)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 1 | 9,2 | 18,3 | 25,8 |
| P-value | 0,712 | 0,678 | 0,692 | 0,813 |

Table 6.9.Final estimates of parameters MA(2)

| Type | Coefficient | t | p |
|---|---|---|---|
| MA 1 | -0,1785 | -3,40 | **0,971** |
| MA 2 | 0,0069 | 0,13 | **0,114** |
| Constant | 5,2423 | 33,39 | 0 |
| SS | **2386,22** | | |
| MS | 6,57 | | |

Table 6.10.Modified Box-Pierce (Ljung-Box) Chi-Square statistic MA(2)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 12,3 | 24,7 | 36,8 | 49,6 |
| P-value | 0,159 | 0,283 | 0,367 | 0,465 |

Probability values of AR(4)-MA(4) combination are higher than 0.05 except AR(2) and the constant term (Table 6.11). According to the test values of Modified Box-Pierce test observed in Table 6.12 there is no evidence of inadequacy of the model. However, the model should be improved to get lower probability values. The next combination that is going to be evaluated is AR(1)-MA(1), as it is found to be significant as well as Lag 4.

Table 6.11.Final estimates of parameters AR(4)-MA(4)

| Type | Coefficient | SE Coefficient | t | p |
|---|---|---|---|---|
| AR 1 | 0,1175 | 1,0186 | 0,12 | **0,908** |
| AR 2 | 0,6048 | 0,2746 | 2,2 | 0,028 |
| AR 3 | 0,3774 | 0,6165 | 0,61 | **0,541** |
| AR 4 | -0,1151 | 0,6586 | -0,17 | **0,861** |
| MA 1 | -0,0369 | 1,012 | -0,04 | **0,971** |
| MA 2 | 0,6553 | 0,4134 | 1,59 | **0,114** |
| MA 3 | 0,5391 | 0,5781 | 0,93 | **0,352** |
| MA 4 | -0,2211 | 0,8004 | -0,28 | **0,783** |
| Constant | 0,034713 | 0,005941 | 5,84 | 0 |
| SS | **115,881** | | | |
| MS | 0,326 | | | |

Table 6.12.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(4)-MA(4)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 1 | 9,2 | 18,3 | 25,8 |
| DF | 3 | 15 | 27 | 39 |
| P-value | 0,79 | 0,869 | 0,894 | 0,949 |

Again, probability values of AR(1)-MA(1) combination are all significantly higher than 0.05 (Table 6.13). As opposed to AR(4)-MA(4) model, the decrease in the probability values of Modified Box-Pierce test statistics is observed in Table

6.14. Normal probability plot of AR(1)-MA(1) model indicates the normality of residuals (Figure 6.8). However, autocorrelogram (Figure 6.9) and partial autocorrelogram (Figure 6.10) have spikes at the 4$^{th}$ lag, which is an evidence of insufficiency of the model. The AR(1)-MA(1) model should be improved.

Table 6.13.Final estimates of parameters AR(1)-MA(1)

| Type | Coefficient | SE Coefficient | t | p |
|---|---|---|---|---|
| AR 1 | -0,0429 | 0,2858 | -0,15 | 0,881 |
| MA 1 | -0,2257 | 0,2783 | -0,81 | 0,418 |
| Constant | 5,4804 | 0,1634 | 33,54 | 0 |
| SS | 2347,28 | | | |
| MS | 6,48 | | | |

Table 6.14.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(1)-MA(1)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 12,2 | 23,1 | 33,9 | 44,2 |
| DF | 9 | 21 | 33 | 45 |
| P-value | 0,203 | 0,337 | 0,426 | 0,505 |

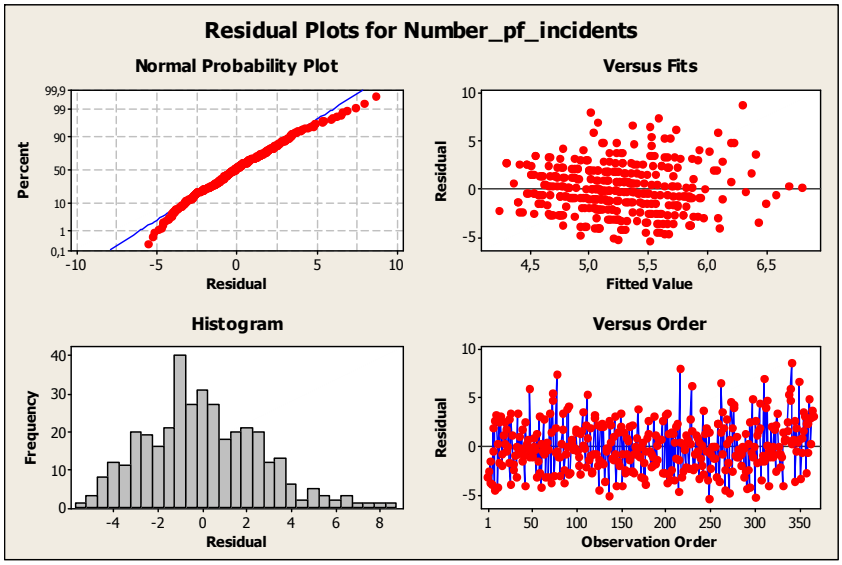


Figure 6.8.Residual plots of AR(1)-MA(1) model

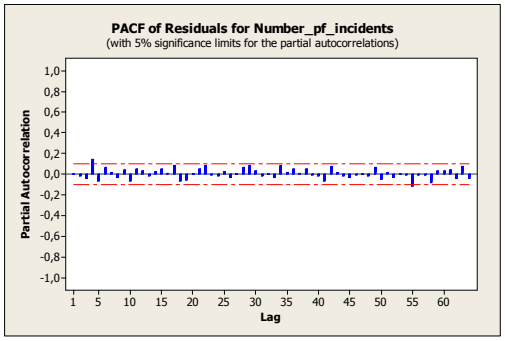


Figure 6.9.Partial autocorrelogram of residuals of AR(1)-MA(1)

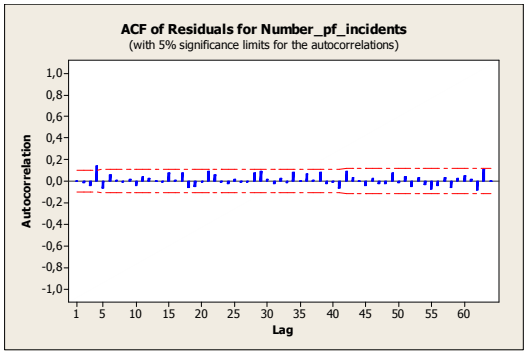


Figure. 6.10.Autocorrelogram of residuals of AR(1)-MA(1)

Then, AR(2)-MA(3) combination is tried to form the model adequately. The probability value results of this combination are much better than prior trials, since only MA(3) (Table 6.15) is insignificant. Removing the MA(3) from the model is necessary to reach the solution. Again there is no problem in residual side in this combination (Table 6.16).

Table 6.15.Final estimates of parameters AR(2)-MA(3)

| Type | Coefficient | SE Coefficient | t | p |
|---|---|---|---|---|
| AR 1 | -1,1665 | 0,1659 | -7,03 | 0,000 |
| AR 2 | -0,6303 | 0,1061 | -5,94 | 0,000 |
| MA 1 | -1,378 | 0,1623 | -8,49 | 0,000 |
| MA 2 | -0,8442 | 0,1311 | -6,44 | 0,000 |
| MA 3 | -0,0154 | 0,05 | -0,31 | 0,759 |
| Constant | 14,6623 | 0,4272 | 34,32 | 0,000 |
| SS | 2287,63 | | | |
| MS | 6,37 | | | |

Table 6.16.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(2)-MA(3)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 2,1 | 11,7 | 22,1 | 30,1 |
| DF | 6 | 18 | 30 | 42 |
| P-value | 0,91 | 0,864 | 0,851 | 0,915 |

Model AR(2)-MA(2) fits to data quiet sufficiently as all the probability values are 0, meaning that all parameters are significant and should be added to the model (Table 6.17). To check the model's adequacy, residuals obtained from the model should be analyzed. At first all p values of Box-pierce statistics are high enough to consider the model adequate (Table 6.18). In addition, plot of residuals is almost normally distributed (Figure 6.11). Also, there are no spikes existing in both of the Autocorrelogram (Figure 6.12) and Partialautocorrelogram (Figure 6.13), meaning no need to improve the model. The last step is forecasting the original and future values of number of incidents in a day.

Table 6.17.Final estimates of parameters AR(2)-MA(2)

| Type | Coefficient | SE Coefficient | t | p |
|---|---|---|---|---|
| AR 1 | -1,14 | 0,1407 | -8,1 | 0,000 |
| AR 2 | -0,6215 | 0,1061 | -6,12 | 0,000 |
| MA 1 | -1,344 | 0,1188 | -11,31 | 0,000 |
| MA 2 | -0,8132 | 0,0821 | -9,9 | 0,000 |
| Constant | 14,5084 | 0,4138 | 35,06 | 0,000 |
| SS | 2257,05 | | | |
| MS | 6,27 | | | |

Table 6.18.Modified Box-Pierce (Ljung-Box) Chi-Square statistic AR(2)-MA(2)

| Lag | 12 | 24 | 36 | 48 |
|---|---|---|---|---|
| Chi Square | 2,1 | 11,6 | 22 | 30,3 |
| DF | 7 | 19 | 31 | 43 |
| P-value | 0,955 | 0,901 | 0,883 | 0,928 |

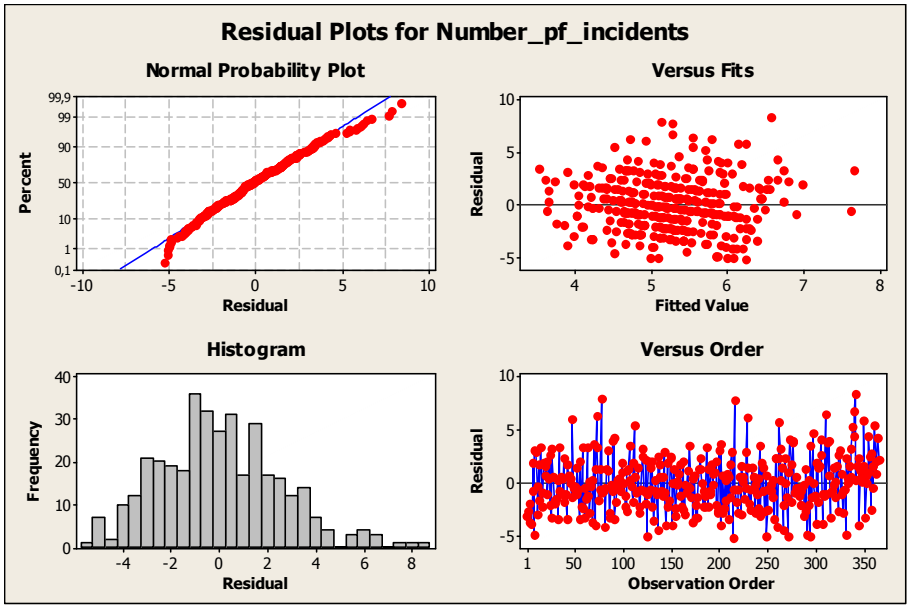


Figure 6.11.Residual plots of AR(2)-MA(2) model

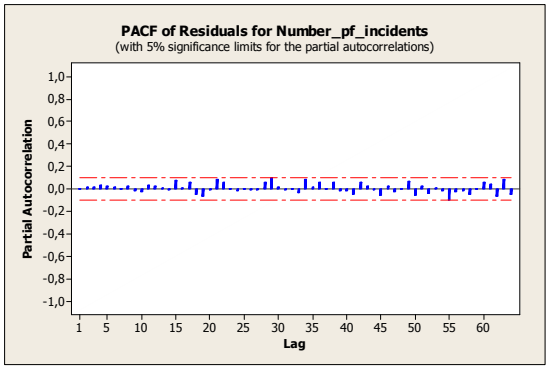


Figure 6.12.Partial autocorrelogram of residuals of AR(2)-MA(2)

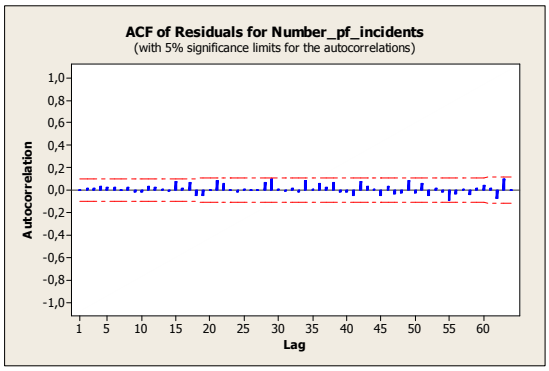


Figure 6.13. Autocorrelogram of residuals of AR(2)-MA(2)

Forecasting results are obtained for all original values and future values. However, the model gives all the future values as 5.23 approximating the mean value, 5.24. An example of forecasted values and their residuals of December 2003 are given in Table 6.19. The mean of the forecasts 5.24 and the standard deviation is 0.036. Predictions are going to be used in spatial disaggregation approach to predict the number of incidents in hot clusters.

Table 6.19.Forecasted values and their residuals of December 2003

| Date | Day | Number of Incidents | Residuals | Forecasted Values |
|---|---|---|---|---|
| 1.December.2003 | Monday | 6 | 1,64 | 4,36 |
| 2.December.2003 | Tuesday | 7 | 0,55 | 6,45 |
| 3.December.2003 | Wednesday | 8 | 3,15 | 4,85 |
| 4.December.2003 | Thursday | 11 | 5,28 | 5,72 |
| 5.December.2003 | Friday | 11 | 4,35 | 6,65 |
| 6.December.2003 | Saturday | 12 | 6,72 | 5,28 |
| 7.December.2003 | Sunday | 15 | 8,41 | 6,59 |
| 8.December.2003 | Monday | 7 | 0,25 | 6,75 |
| 9.December.2003 | Tuesday | 6 | 1,59 | 4,41 |
| 10.December.2003 | Wednesday | 8 | 2,35 | 5,65 |
| 11.December.2003 | Thursday | 2 | -4,09 | 6,09 |
| 12.December.2003 | Friday | 4 | 0,35 | 3,65 |
| 13.December.2003 | Saturday | 7 | 1,19 | 5,81 |
| 14.December.2003 | Sunday | 6 | 0,12 | 5,88 |
| 15.December.2003 | Monday | 6 | 1,57 | 4,43 |
| 16.December.2003 | Tuesday | 12 | 5,86 | 6,14 |
| 17.December.2003 | Wednesday | 3 | -3,24 | 6,24 |
| 18.December.2003 | Thursday | 4 | -0,04 | 4,04 |
| 19.December.2003 | Friday | 6 | 0,65 | 5,35 |
| 20.December.2003 | Saturday | 7 | 1,03 | 5,97 |
| 21.December.2003 | Sunday | 9 | 4,32 | 4,68 |
| 22.December.2003 | Monday | 9 | 2,46 | 6,54 |
| 23.December.2003 | Tuesday | 3 | -2,47 | 5,47 |
| 24.December.2003 | Wednesday | 7 | 2,83 | 4,17 |
| 25.December.2003 | Thursday | 8 | 1,57 | 6,43 |
| 26.December.2003 | Friday | 5 | -0,42 | 5,42 |
| 27.December.2003 | Saturday | 10 | 5,47 | 4,53 |
| 28.December.2003 | Sunday | 9 | 2,01 | 6,99 |
| 29.December.2003 | Monday | 6 | 0,83 | 5,17 |
| 30.December.2003 | Tuesday | 9 | 4,18 | 4,82 |
| 31.December.2003 | Wednesday | 9 | 2,20 | 6,80 |

## 6.2. Simple spatial disaggregation approach (SSDA) of spatio-temporal crime prediction model

The aim of SSDA is to produce cluster forecasts that give minimized forecast errors. It is a kind of spatio-temporal forecasting technique, explores and establishes weekday-specific clusters of time (Al-Madfai et.al.,2006). Hence, the clusters are allowed to be different than each other. Daily specific crime incidents are aggregated and than appropriate clusters are formed by STAC hot clusters for both types of distance metrics. The advantages of STAC are explained in the third chapter. Of course it is not always possible to make the same pattern for all days of year every time. Other effects like a special event, weather and any factors may happen but these things are ignored in this study.

Using the time series model and forecasts, the first thing is to assign each forecasted value to the clusters. The number of clusters of each day differs as seen in Table 6.20 and Table 6.21.

Table 6.20.Number of incidents for STAC with Manhattan distance metric clusters per day.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Total # of Incidents in Clusters | Total # of Incidents | % of Incidents found in Clusters |
|---|---|---|---|---|---|---|---|---|
| Monday | 11 | 19 | 8 | 16 | 95 | 149 | 281 | 53 |
| Tuesday | 36 | 26 | 77 | 0 | 0 | 139 | 279 | 49 |
| Wednesday | 13 | 8 | 24 | 9 | 72 | 126 | 262 | 48 |
| Thursday | 18 | 20 | 13 | 88 | 0 | 139 | 249 | 56 |
| Friday | 37 | 19 | 96 | 0 | 0 | 152 | 286 | 53 |
| Saturday | 24 | 142 | 0 | 0 | 0 | 166 | 290 | 57 |
| Sunday | 8 | 12 | 104 | 0 | 0 | 120 | 238 | 50 |

Table 6.21.Number of incidents for STAC with Euclidean distance metric clusters per day.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 | Total # of Incidents | Total # of | % of Incidents found |
|---|---|---|---|---|---|---|---|---|---|---|
| Monday | 11 | 8 | 11 | 14 | 58 | 8 | 23 | 133 | 281 | 0.47 |
| Tuesday | 37 | 22 | 8 | 20 | 36 | 14 | | 137 | 279 | 0,49 |
| Wednesday | 10 | 9 | 18 | 71 | | | | 108 | 262 | 0,41 |
| Thursday | 12 | 13 | 37 | 8 | 13 | | | 83 | 249 | 0,33 |
| Friday | 28 | 11 | 25 | 21 | 8 | 45 | | 138 | 286 | 0,48 |
| Saturday | 8 | 9 | 13 | 100 | 21 | | | 151 | 290 | 0,52 |
| Sunday | 9 | 9 | 11 | 70 | | | | 99 | 238 | 0,42 |

Assignment is done according to the formulation;

$$O_{tj} = B_{tj} * y_t, \text{ with } \sum_{\forall j} B_{tj} = 1 \qquad \text{Eq. (6.1)}$$

$O_{tj}$ represent the forecast of cluster j on day t, $y_t$ is the forecast of day t obtained from the model. $B_{tj}$ is the spatial forecast disaggregation weights allocated to each cluster per day. There are several ways of determining the spatial forecast disaggragetion weights but here another method is applied as it is giving more general sight to the model. Another reason is not to have enough data to get sufficient results with the other methods indicated in Chapter 3. In this method, spatial forecast disaggragetion (SFD) weights are set equal by counting observations which happened on each day of the week in the area of a cluster (Table 6.20, 6.21) and then percentage of incidents in clusters to total number of incidents in all the clusters are calculated. Therefore, the weights for each cluster per day are assigned in Table 6.22 and Table 6.23. Firstly, total number of incidents is counted (Table 6.20, 6.21) and weights are found (Table 6.22, 6.23). Also, in Table 6.12 and Table 6.13 percentage of incidents represented by clusters are calculated. Approximately, %51 of total incidents is represented by STAC with Euclidean distance metric clusters. However, representation percentage is smaller (% 45) in STAC with Manhattan distance metric clusters. This is expected as there is 0.4 $km^2$ difference between the total areas of clusters.

Table 6.22.SFD weights assigned to each STAC with Euclidean distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| Monday | 0,074 | 0,128 | 0,054 | 0,107 | 0,638 |
| Tuesday | 0,259 | 0,187 | 0,554 | | |
| Wednesday | 0,103 | 0,063 | 0,190 | 0,071 | 0,571 |
| Thursday | 0,129 | 0,144 | 0,094 | 0,633 | 0,000 |
| Friday | 0,243 | 0,125 | 0,632 | | |
| Saturday | 0,145 | 0,855 | | | |
| Sunday | 0,067 | 0,100 | 0,867 | | |

Table 6.23.SFD weights assigned to each STAC with Manhattan distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 |
|---|---|---|---|---|---|---|---|
| Monday | 0,083 | 0,060 | 0,083 | 0,105 | 0,436 | 0,060 | 0,173 |
| Tuesday | 0,270 | 0,161 | 0,058 | 0,146 | 0,263 | 0,102 | |
| Wednesday | 0,093 | 0,083 | 0,167 | 0,657 | | | |
| Thursday | 0,145 | 0,157 | 0,446 | 0,096 | 0,157 | | |
| Friday | 0,203 | 0,080 | 0,181 | 0,152 | 0,058 | 0,326 | |
| Saturday | 0,053 | 0,060 | 0,086 | 0,662 | 0,139 | | |
| Sunday | 0,091 | 0,091 | 0,111 | 0,707 | | | |

Then $O_{tj}$ values are calculated for each cluster of the year 2003. For example for STAC with Euclidean distance metric clusters;

For the 1st of January,

$O_{1,1}$ = 0.103(Wednesday-1st cluster) * 5.17(Number of crime predicted for the 1st of January)

$O_{1,1}$ = 0.533

All the cluster forecasts are evaluated according to the weights and predictions per cluster per day. “0” indicates there is no such cluster exists in that day. Table 6.24 indicates the cluster forecasts of January for STAC with Euclidean distance metric clusters.

Table 6.24.Cluster forecasts for January.

| Date | Day | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|---|
| 1.January.2003 | Wednesday | 0,533 | 0,328 | 0,984 | 0,369 | 2,952 |
| 2.January.2003 | Thursday | 0,615 | 0,683 | 0,444 | 3,007 | 0,000 |
| 3.January.2003 | Friday | 1,134 | 0,582 | 2,942 | 0,000 | 0,000 |
| 4.January.2003 | Saturday | 0,710 | 4,202 | 0,000 | 0,000 | 0,000 |
| 5.January.2003 | Sunday | 0,324 | 0,486 | 4,210 | 0,000 | 0,000 |
| 6.January.2003 | Monday | 0,348 | 0,600 | 0,253 | 0,506 | 3,002 |
| 7.January.2003 | Tuesday | 1,341 | 0,969 | 2,868 | 0,000 | 0,000 |
| 8.January.2003 | Wednesday | 0,607 | 0,373 | 1,120 | 0,420 | 3,360 |
| 9.January.2003 | Thursday | 0,506 | 0,562 | 0,365 | 2,472 | 0,000 |
| 10.January.2003 | Friday | 1,475 | 0,757 | 3,826 | 0,000 | 0,000 |
| 11.January.2003 | Saturday | 0,918 | 5,432 | 0,000 | 0,000 | 0,000 |
| 12.January.2003 | Sunday | 0,266 | 0,399 | 3,456 | 0,000 | 0,000 |
| 13.January.2003 | Monday | 0,392 | 0,678 | 0,285 | 0,571 | 3,389 |
| 14.January.2003 | Tuesday | 1,472 | 1,063 | 3,148 | 0,000 | 0,000 |
| 15.January.2003 | Wednesday | 0,483 | 0,297 | 0,891 | 0,334 | 2,674 |
| 16.January.2003 | Thursday | 0,773 | 0,859 | 0,558 | 3,780 | 0,000 |
| 17.January.2003 | Friday | 1,268 | 0,651 | 3,291 | 0,000 | 0,000 |
| 18.January.2003 | Saturday | 0,605 | 3,579 | 0,000 | 0,000 | 0,000 |
| 19.January.2003 | Sunday | 0,427 | 0,641 | 5,556 | 0,000 | 0,000 |
| 20.January.2003 | Monday | 0,387 | 0,668 | 0,281 | 0,563 | 3,341 |
| 21.January.2003 | Tuesday | 1,296 | 0,936 | 2,771 | 0,000 | 0,000 |
| 22.January.2003 | Wednesday | 0,628 | 0,386 | 1,159 | 0,434 | 3,476 |
| 23.January.2003 | Thursday | 0,596 | 0,662 | 0,430 | 2,912 | 0,000 |
| 24.January.2003 | Friday | 1,298 | 0,667 | 3,368 | 0,000 | 0,000 |
| 25.January.2003 | Saturday | 0,891 | 5,272 | 0,000 | 0,000 | 0,000 |
| 26.January.2003 | Sunday | 0,281 | 0,422 | 3,653 | 0,000 | 0,000 |
| 27.January.2003 | Monday | 0,352 | 0,607 | 0,256 | 0,511 | 3,037 |
| 28.January.2003 | Tuesday | 1,671 | 1,207 | 3,574 | 0,000 | 0,000 |
| 29.January.2003 | Wednesday | 0,506 | 0,311 | 0,934 | 0,350 | 2,803 |
| 30.January.2003 | Thursday | 0,596 | 0,662 | 0,430 | 2,914 | 0,000 |
| 31.January.2003 | Friday | 1,332 | 0,684 | 3,456 | 0,000 | 0,000 |

The number of future crime incidents is found 5,23 with Box-Jenkins ARIMA model. The last step is to assign the forecasted value with respect to weights to each cluster. Assigned values are indicated at Table 6.25 and 6.26. Values should be rounded up and down as crime incident number is an integer value indicated at Table 6.27 and Table 6.28.

Spatio-temporal crime prediction model is applied to both types of distance metric however, to test which distance metric is fitting better to the model, the next step is to evaluate spatio-temporal forecast errors of SDA. Spatio-temporal mean root square estimate is used to detect daily deviation of the model. The main advantage of this measure resides in its easy implementation and in fact to its widely understood temporal counterparts (Al-Madfai et al., 2006).

The formulation of STMRSE is;

$$STRMSE(\varepsilon) = \sqrt{\frac{1}{n}\sum_{t}^{n}\sum_{j}^{m_t}\frac{(Observed_{tj} - O_{tj})^2}{m_t}} \quad \text{Eq. (6.2.2)}$$

Where n being the total number of days (Al-Madfai et al., 2006).

STRMSE is calculated with counting all the observations per clusters per day and found as 1,48 and 1,08 for STAC-Euclidean and STAC-Manhattan, respectively. Manhattan distance gives better result as the value is nearer to "0" also both values are not big when concerning crime incidents. Also, STRMSE of all the clusters are calculated to define the accuracy for each cluster. Generally daily STRMSE's of clusters are smaller in STAC-Manhattan than Euclidean where sixth cluster of Monday has only 0.390 STRMSE.

Table 6.25.Forecasted values for each STAC with Euclidean distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| Monday | 0,386 | 0,669 | 0,282 | 0,560 | 3,337 |
| Tuesday | 1,355 | 0,978 | 2,897 | | |
| Wednesday | 0,540 | 0,329 | 0,994 | 0,371 | 2,986 |
| Thursday | 0,677 | 0,753 | 0,492 | 3,311 | |
| Friday | 1,273 | 0,654 | 3,305 | | |
| Saturday | 0,756 | 4,472 | | | |
| Sunday | 0,349 | 0,523 | 4,534 | | |

Table 6.26.Rounded forecasted values for each STAC with Euclidean distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| Monday | 0-1 | 0-1 | 0-1 | 0-1 | 3-4 |
| Tuesday | 1-2 | 0-1 | 2-3 | | |
| Wednesday | 0-1 | 0-1 | 0-1 | 0-1 | 2-3 |
| Thursday | 0-1 | 0-1 | 0-1 | 3-4 | |
| Friday | 1-2 | 0-1 | 3-4 | | |
| Saturday | 0-1 | 4-5 | | | |
| Sunday | 0-1 | 0-1 | 4-5 | | |

Table 6.27.Forecasted values for each STAC with Manhattan distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 |
|---|---|---|---|---|---|---|---|
| Monday | 0,433 | 0,315 | 0,433 | 0,551 | 2,281 | 0,315 | 0,904 |
| Tuesday | 1,412 | 0,840 | 0,305 | 0,764 | 1,374 | 0,534 | |
| Wednesday | 0,484 | 0,436 | 0,872 | 3,438 | | | |
| Thursday | 0,756 | 0,819 | 2,331 | 0,504 | 0,819 | | |
| Friday | 1,061 | 0,417 | 0,947 | 0,796 | 0,303 | 1,705 | |
| Saturday | 0,277 | 0,312 | 0,450 | 3,464 | 0,727 | | |
| Sunday | 0,475 | 0,475 | 0,581 | 3,698 | | | |

Table 6.28.Rounded forecasted values for each STAC with Manhattan distance metric clusters.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 |
|---|---|---|---|---|---|---|---|
| Monday | 0-1 | 0-1 | 0-1 | 0-1 | 2-3 | 0-1 | 0-1 |
| Tuesday | 1-2 | 0-1 | 0-1 | 0-1 | 1-2 | 0-1 | |
| Wednesday | 0-1 | 0-1 | 0-1 | 3-4 | | | |
| Thursday | 0-1 | 0-1 | 2-3 | 0-1 | 0-1 | | |
| Friday | 1-2 | 0-1 | 0-1 | 0-1 | 0-1 | 1-2 | |
| Saturday | 0-1 | 0-1 | 0-1 | 3-4 | 0-1 | | |
| Sunday | 0-1 | 0-1 | 0-1 | 3-4 | | | |

Forecasted value of 5.23 is assigned to the clusters according to their weights and resulting numbers are rounded as the numbers should be integer. According to the Tables 6.25 and 6.26, on Monday at cluster 1, there will be 0 or 1 crime incidents happened at both type of clusters. To give another example on Wednesday at cluster 5 the range of crime incidents is between 2 and 3 for STAC with Euclidean

distance metric algorithm. To demonstrate these values on clusters, STAC hot clusters for each day are mapped. Numbers on clusters represent the forecasted values for that cluster.

Cluster 1 is located on commercial areas and an empty market area in Bişkek Street illustrated at Figure 6.14. It is the most accurate and best fitting cluster having smaller STRMSE than other clusters. Second cluster covers relatively big area Kazakistan Street passing. Area is both commercial and residential. In the area that the third and the fourth clusters involve, burglary and auto related crimes happened supporting the land-use area which is residential. The last cluster dispersed into the commercial area in Çankaya Region that is more probable to expose to a crime activity.

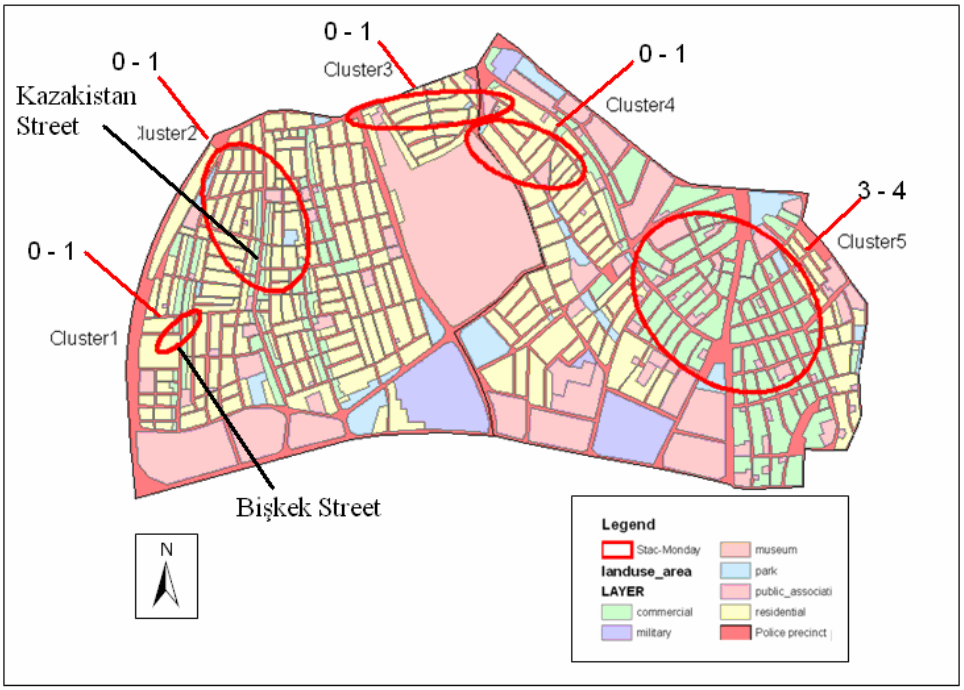



Figure 6.14.STAC with Euclidean distance metric hot clusters for Monday.

There are three clusters in Tuesday on Figure 6.15, where the first one covers the most active area including Seventh Street in Bahçelievler. Also, police officer in

Bahçelievler police station told that Seventh Street is the most attractive area giving opportunity for offenders in Bahçelievler. Both commercial and residential area is present in that place. Another interesting point is, only clusters in Bahçelievler on Tuesday and Friday have more than one criminal event forecasted. Second cluster again contains Beşevler region and the third cluster located in commercial areas in Çankaya.

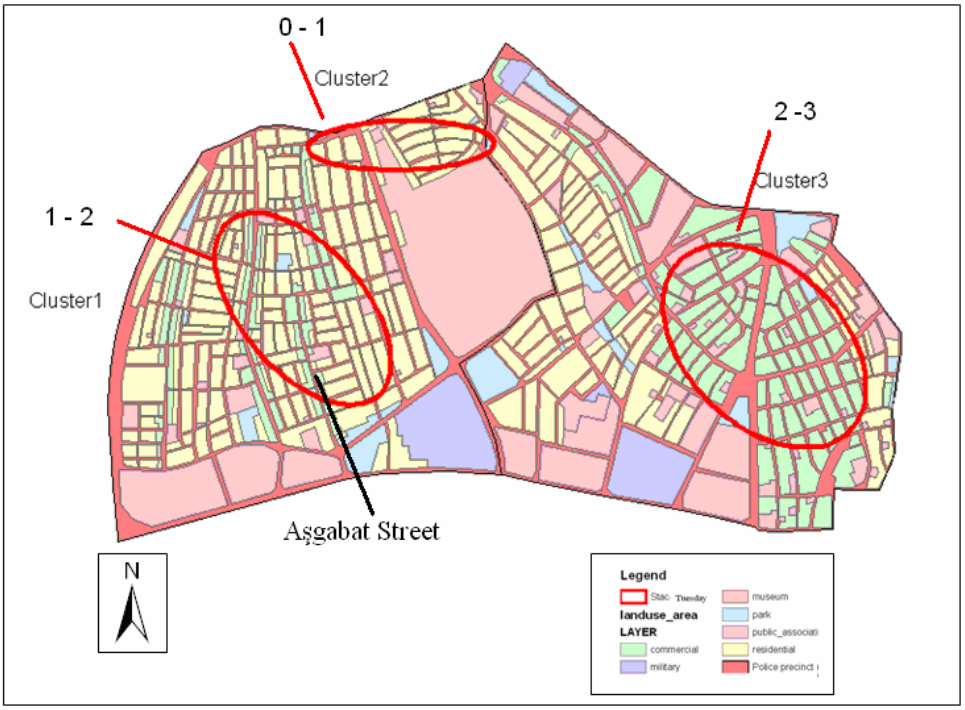



Figure 6.15.STAC with Euclidean distance metric hot clusters for Tuesday.

Clusters of Wednesday are mapped and given in Figure 6.16. The first cluster of Wednesday resides on area near Bişkek Street. Second cluster is small and so, more significant cluster located on the area where 6$^{th}$ and 7$^{th}$ street intersect. Forecasting crime in a smaller area means better prediction as numbers are giving information about a more specific place. The third cluster covers relatively larger area in Bahçelievler. The fourth cluster is again located specifically on Şerefli Street. The least fitting cluster is the fifth cluster located in Çankaya.

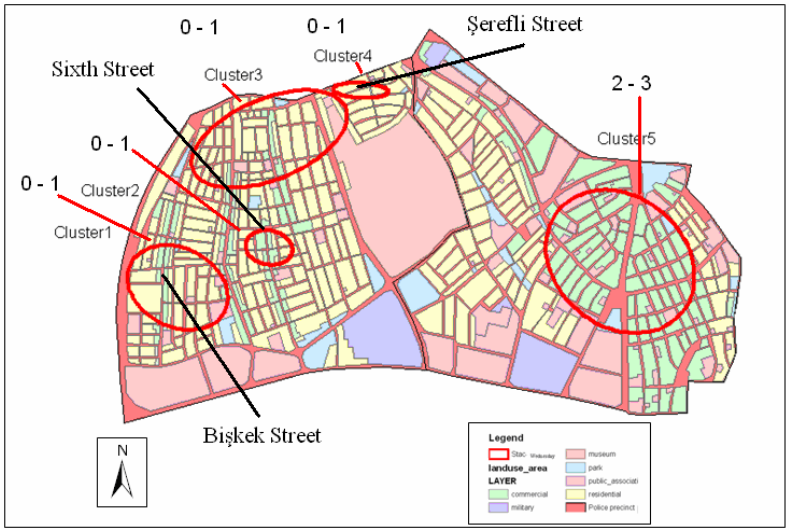


Figure 6.16.STAC with Euclidean distance metric hot clusters for Wednesday.

The first cluster of Thursday includes the area between Kazakistan and Taşkent Streets. Also, second clusters reside in a similar mixed area in Bahçelievler. The third cluster is relatively small cluster covering an area with a school. This time different from the previous days, last cluster is elongated to the north including the area near the Ministry of Health.

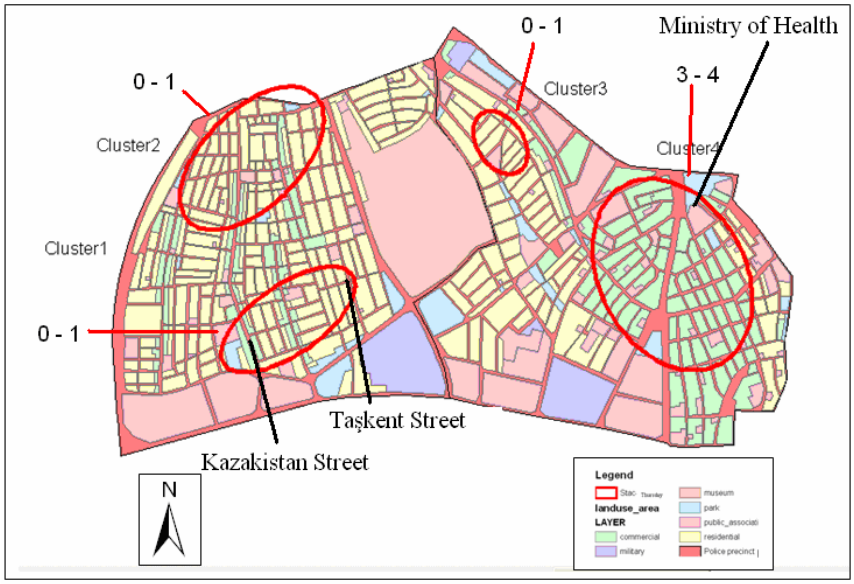


Figure 6.17.STAC with Euclidean distance metric hot clusters for Thursday.

It can be seen from maps in Figure 6.18, 6.19, and 6.20 that the number of clusters decreasing when it comes to the end of week. Friday has three clusters where the first one located in area including Aşgabat and Kazakistan Streets. The second cluster is important here as it is the smallest in size. It is located towards the Bahri Üçok Street having residential areas. Last cluster is elongated to Maltepe part,which is different from the other day's configurations.

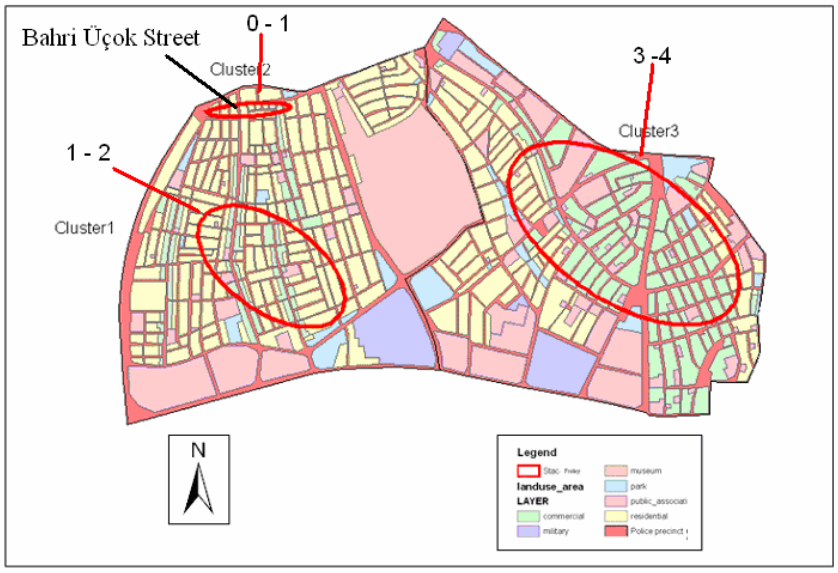



Figure 6.18.STAC with Euclidean distance metric hot clusters for Friday.

Saturday has only two clusters representing the two sides of police precincts (Figure 6.19). The cluster in Bahçelievler includes almost all of the Aşgabat Street and the second include all the commercial area in Çankaya. Also, number of incidents forecasted for clusters in Çankaya region increases on weekend. For Saturday, the reason would be the crowd in commercial areas and for Sunday empty commercial areas due to holiday.

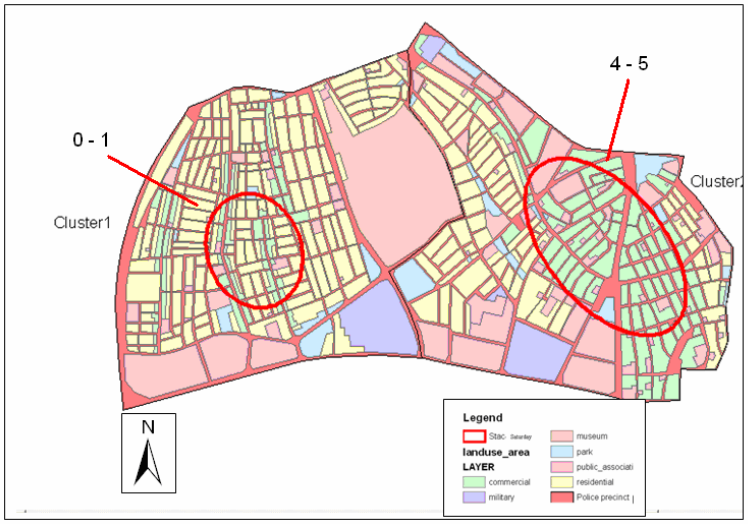


Figure 6.19.STAC with Euclidean distance metric hot clusters for Saturday.

Sunday has two clusters in Bahçelievler illustrated at Figure 6.20. The first one has small area between Aşgabat and Kazakistan Streets. The second one is at the middle of Emek and highlighted for the first time at the last day of the week. The interesting point is the most of the crime incidents are burglary in spite of Sunday. The last cluster residing at the same location has more number of incidents forecasted.

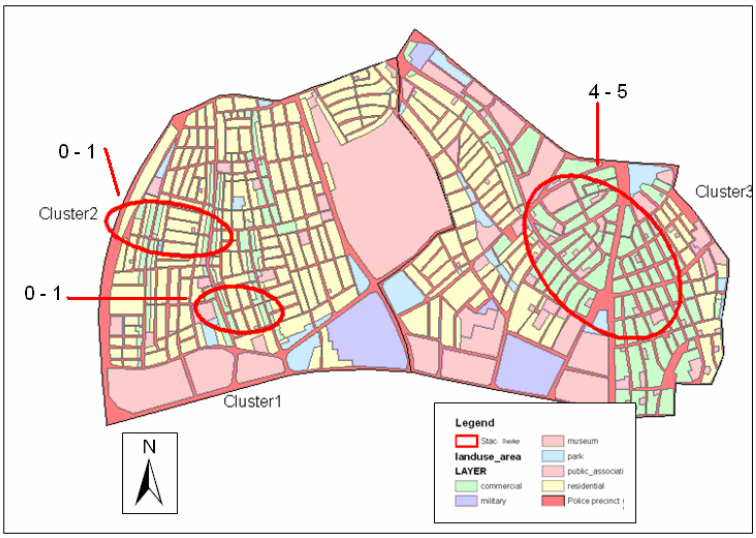


Figure 6.20.STAC with Euclidean distance metric hot clusters for Sunday.

The second part is STAC with Manhattan distance metric part and at the first glance number of clusters increase in this part. The reason is explained in the previous section when giving information about the structure of clusters. Clusters on Monday is illustrated in Figure 6.21. Number of clusters on Monday is consistent with STAC with Euclidean distance metric as both have the bigger number of clusters. The region of the first cluster is familiar as the area is included in the clusters. The second cluster includes 79$^{th}$ Street in Emek. The third cluster is located in Maltepe elongated to north-south direction. The fourth cluster is small but covers Maltepe Bazaar. The fifth cluster is located at west side of the Atatürk Avenue, where the last two are located in Ziya Gökalp Avenue and Konur Street, respectively.

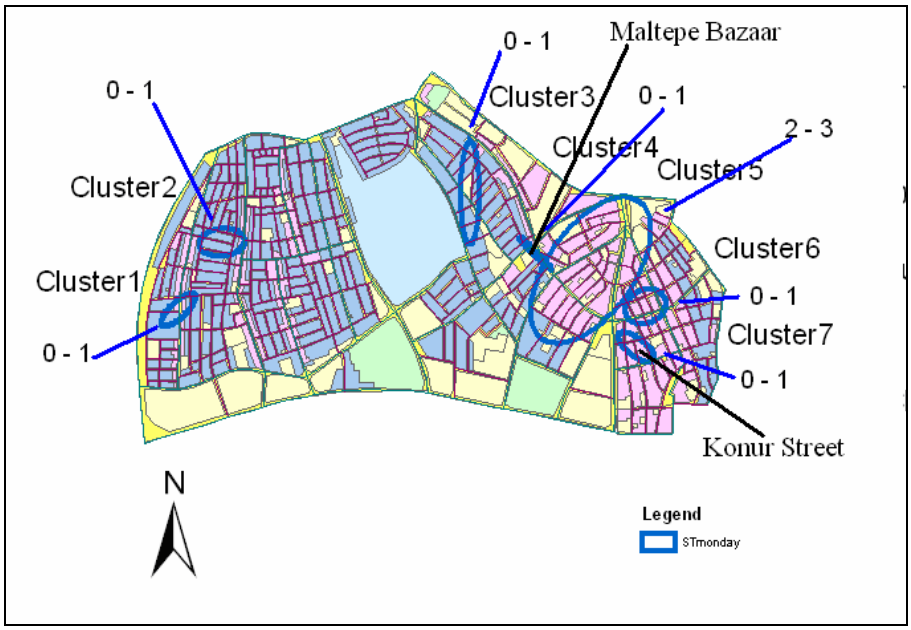


Figure 6.21.STAC with Manhattan distance metric hot clusters for Monday.

To look at the clusters in Figure 6.22, Tuesday has a bigger cluster in Bahçelievler including Aşgabat, Kazakistan and Azerbaycan Streets. The second cluster is located in Beşevler towards the university campuas area and the third one is located in again near Maltepe Bazaar. Note that, Maltepe Bazaar does not exist at

the same area today but in year 2003 it was. The fourth cluster covers Belediye Hospital and Ministry of Health. The area is mainly commercial although some public associations exist. The fifth one covers the area between Atatürk Avenue and Mithatpaşa Street and the last cluster is located in mixed area around Kocatepe Mosque.

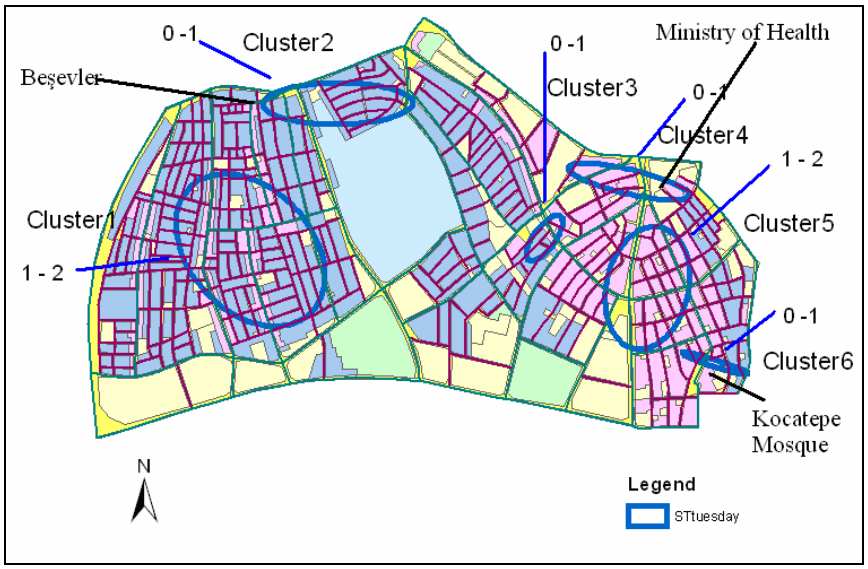


Figure 6.22. STAC with Manhattan distance metric hot clusters for Tuesday.

The first cluster of Wednesday again covers the commercial areas on Bişkek Street as seen in Figure 6.23. The Second cluster is small and has no significance. The third cluster covers the area near Ankaray subway stop and the last cluster is located in Merkez Çankaya police precinct part. Wednesday is different than the other days for STAC with Manhattan distance metric as there is only one cluster representing Çankaya region. Generally in other days, there are more than two clusters in Çankaya.

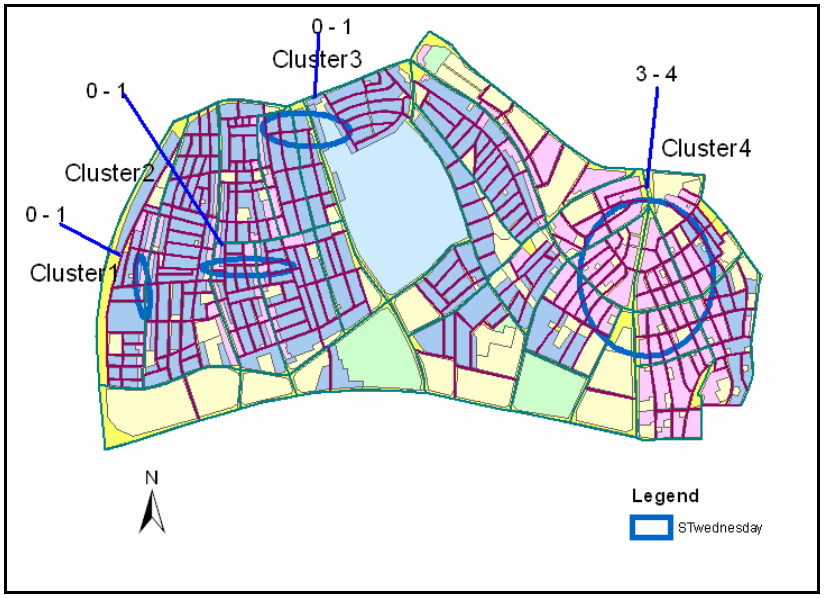


Figure 6.23.STAC with Manhattan distance metric hot clusters for Wednesday.

Among the clusters of Thursday (Figure 6.24) the first and the fourth clusters indicate different areas than observed before. Cluster one includes the area at two side of Kazakistan Street in Bahçelievler. In addition, it is the first time that a cluster covers the residential areas in Çankaya including Kurtuluş Park.

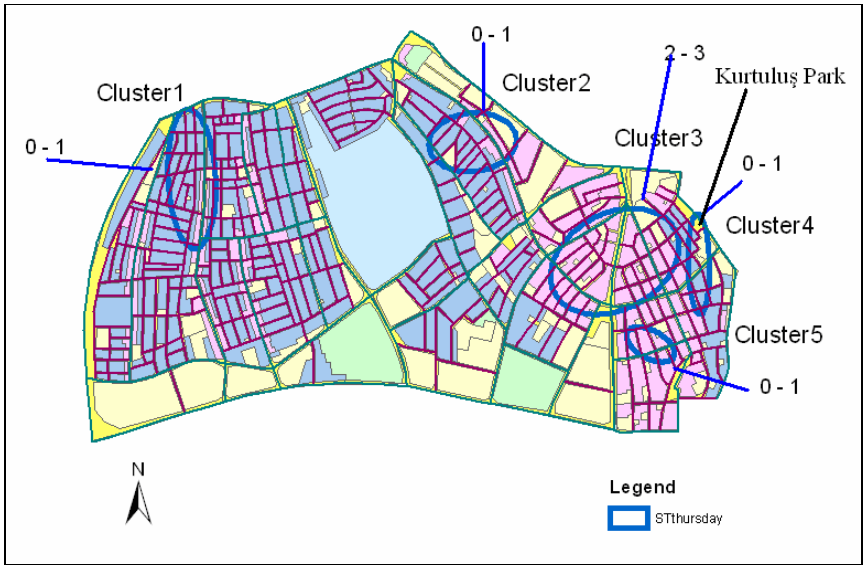


Figure 6.24.STAC with Manhattan distance metric hot clusters for Thursday.

The difference of Friday hot clusters (Figure 6.25) is the area covered in Çankaya region. The area is smaller than all the other days. Area is represented by three small clusters including mostly east side of Atatürk Avenue.

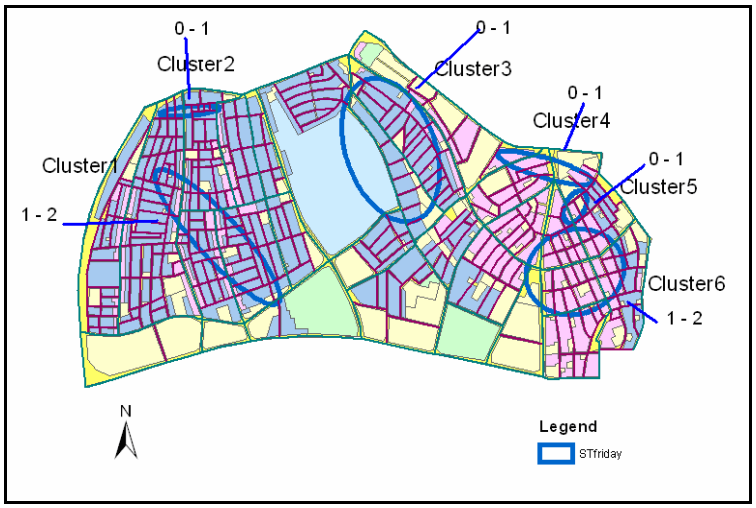


Figure 6.25. STAC with Manhattan distance metric hot clusters for Friday.

The first cluster of Saturday includes almost all Aşgabat Street and the second has a boundry with 5$^{th}$ and 6$^{th}$ Streets. The last cluster has only one street segment which is Konur Street illustrated in Figure 6.26.

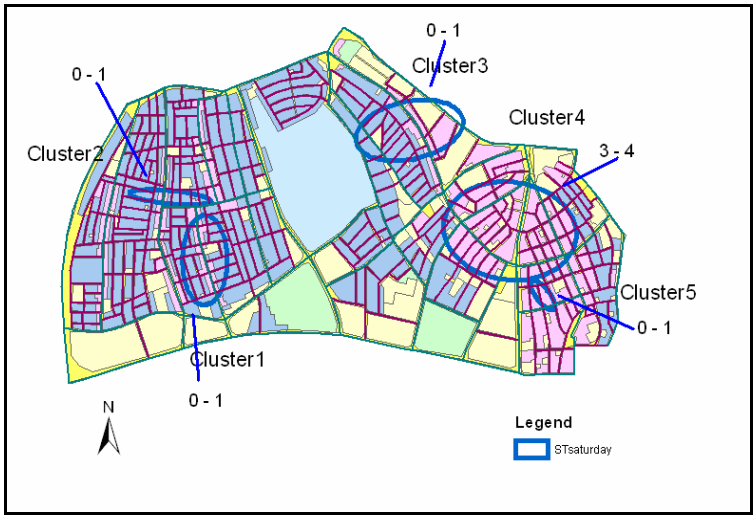


Figure 6.26.STAC with Manhattan distance metric hot clusters for Saturday.

Clusters of Sunday are similar to clusters exist in other days. The difference of Sunday is that it has clusters mostly in southern part of the study area.

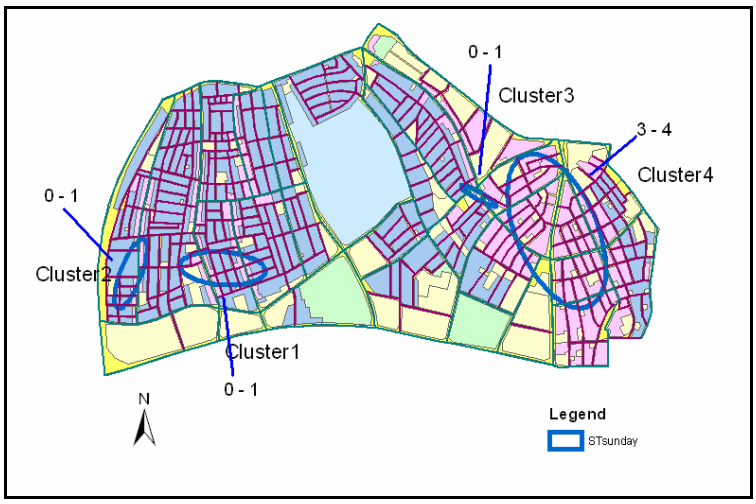


Figure 6.27.STAC with Manhattan distance metric hot clusters for Sunday.

### 6.3. Model validation

To validate the model last seven days are separated from the yearly data and the probable number of crime incidents is predicted. Last seven days are used for model validation because there is no future value is available. Box-Jenkins ARIMA model is applied and the forecasted values are found 5,19 for each day. Forecasted values are again assigned to each cluster according to the SFD weights found earlier in the chapter. The results are indicated in Table 23 and 24.

Table 6.29. Number of incidents predicted for model validation for Euclidean distance metric

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| Monday | 0,383 | 0,662 | 0,279 | 0,557 | 3,309 |
| Tuesday | 1,344 | 0,971 | 2,875 | 0,000 | 0,000 |
| Wednesday | 0,535 | 0,330 | 0,989 | 0,371 | 2,966 |
| Thursday | 0,672 | 0,747 | 0,485 | 3,286 | 0,000 |
| Friday | 1,263 | 0,649 | 3,278 | 0,000 | 0,000 |
| Saturday | 0,750 | 4,440 | 0,000 | 0,000 | 0,000 |
| Sunday | 0,346 | 0,519 | 4,498 | 0,000 | 0,000 |

Table 6.30. Number of incidents predicted for model validation for Manhattan distance metric

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 |
|---|---|---|---|---|---|---|---|
| Monday | 0,429 | 0,312 | 0,429 | 0,546 | 2,263 | 0,312 | 0,898 |
| Tuesday | 1,402 | 0,833 | 0,303 | 0,758 | 1,364 | 0,530 | 0,000 |
| Wednesday | 0,481 | 0,433 | 0,865 | 3,412 | 0,000 | 0,000 | 0,000 |
| Thursday | 0,750 | 0,813 | 2,314 | 0,500 | 0,813 | 0,000 | 0,000 |
| Friday | 1,053 | 0,414 | 0,940 | 0,790 | 0,301 | 1,692 | 0,000 |
| Saturday | 0,275 | 0,309 | 0,447 | 3,437 | 0,722 | 0,000 | 0,000 |
| Sunday | 0,472 | 0,472 | 0,577 | 3,670 | 0,000 | 0,000 | 0,000 |

To compare the validity of Euclidean and Manhattan distance metric, both the predicted and the observed number of incidents are revealed in Tables 6.29, 6.30, 6.31, 6.32. When the observed and predicted numbers of incidents are compared, Euclidean distance metric is found to be more accurate. In Euclidean distance metric, the difference between the observed and the predicted value is bigger than 1 in only three clusters, whereas it is four in Manhattan distance metric. Also, the number of clusters which have difference between the predicted and the observed number of incidents smaller than 0,1 is three in Euclidean, whereas it is zero in Manhattan distance metrics. It is observed that validity of the smaller clusters is more than the bigger ones. Bigger clusters are generally located in Merkez Çankaya police precinct where the number of pickpocketing is higher than the other crime types.

Table 6.31. Number of incidents observed in the clusters for Euclidean distance metric

| Date | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 |
|---|---|---|---|---|---|
| 25.December-Thursday | 1 | 0 | 0 | 0 | |
| 26.December-Friday | 2 | 0 | 2 | | |
| 27.December-Saturday | 1 | 2 | | | |
| 28.December-Sunday | 0 | 0 | 2 | | |
| 29.December-Monday | 0 | 1 | 0 | 0 | 3 |
| 30.December-Tuesday | 1 | 1 | 2 | | |
| 31.December-Wednesday | 0 | 0 | 1 | 0 | 3 |

Table 6.32. Number of incidents observed in the clusters for Manhattan distance metric

| Date | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 |
|---|---|---|---|---|---|---|---|
| 25.December | 1 | 0 | 1 | 0 | 1 | | |
| 26.December | 2 | 0 | 0 | 1 | 0 | 1 | |
| 27.December | 0 | 1 | 1 | 2 | 1 | | |
| 28.December | 0 | 0 | 0 | 2 | | | |
| 29.December | 0 | 0 | 0 | 1 | 2 | 0 | 0 |
| 30.December | 1 | 0 | 0 | 0 | 1 | 1 | |
| 31.December | 0 | 0 | 0 | 2 | | | |

# CHAPTER 7

# DISCUSSION AND CONCLUSION

## 7.1.Discussion of the clustering algorithms

The first part of the methodology of this thesis is to generate clusters according to different approaches and compare the clusters with respect to land-use, algorithms, covered area, and suitability to a spatio-temporal crime prediction model. In order to determine the most suitable clustering algorithm, K-means, Nnh hierarchical, spatio-temporal analysis of crime (STAC), fuzzy, ISODATA, and geographical analysis machine (GAM) clustering techniques are implemented and examined. All clustering algorithms generate clusters which are located in different sizes, numbers and orientation. These differences influence the decision of choosing the right model. One of the most important selection criteria is the capability of reflecting real dense areas not by including all the area. This is important for crime prediction model, as model is generated for security dealers and should be meaningful for crime prevention.

To sum up,

- Clusters of K-means, fuzzy and ISODATA (optimization/partitioning based algorithms) cover all the observations in the study area. Hence, it is difficult to detect dense crime areas with these algorithms. Also, spatial outliers are forced to include in clusters that changes the orientation of clusters. One advantage of these clustering algorithms is to let user to define the number of clusters in the study area.

- Nearest neighbor Hierarchical approach is applied to the data. However, results are not found to be meaninful because the clusters are too much and small. Hence, It is not effective for crime prediction model as the data on hand can not handle that much clusters.

- GAM has different working principle than the other clustering algorithms in the study. GAM uses the underlying population in the area to form clusters in terms of a defined variable. Number of crimes with population of neighborhoods are analyzed by GAM. Clusters generated in Eti, Korkut Reis, Sağlık, Kızılay, Cumhuriyet, Fidanlık, Kocatepe and Meşrutiyet neighborhoods, which have smaller surface area than the other neighborhoods.

- STAC is selected to use in spatio-temporal crime prediction model. STAC has several advantages over other clustering algorithms. STAC shows relatively dense crime areas than the other clustering algorithms which is significant in crime prediction model.

- The structure of STAC and Nnh Hierarchical clustering algorithms are similar that they have generally larger clusters at Merkez Çankaya Police precincts. This is an advantage in crime prediction model. As land use in Merkez Çankaya police precinct is more homogeneous, crime incidents are belong to the same type of crime. For example, pickpocketing is dominant in Çankaya especially in Kızılay Square.

- STAC with 300 m fixed distance and 10 minimum number of incidents per cluster is chosen which covers %87 of observations. This means that STAC clusters with defined constarints are able to cover most of the observations without covering all the area.

### 7.2.Discussion of the distance metrics

The applied distance metrics give different configuration to clusters. However, all the clustering methods can not be interpreted by different distance metrics. This can be because of the algorithm or the limitation of related softwares. The next

step is to evaluate the selected algorithm "STAC" based on different distance metrics. Both Manhattan and Euclidean distance metrics are applied with STAC to form clusters to all the data. The structure and effect of distance metrics are compared and evaluated to form background for crime prediction model. STAC Manhattan and Euclidean mostly differs according to the orientation of clusters. In Manhattan, clusters elongated in north-south direction with decreasing in area, whereas in Euclidean clusters are mostly circular. It is generally covers the major streets in the same direction with clusters. Some clusters in STAC-Manhattan are so small not to indentify meaningful areas.Area coverd by Euclidean provides larger clusters, which supply better control areas for police.

To sum up;

- Manhattan distance configuration is nearer to original network.
- STAC clusters with Manhattan distance represents denser crime area than STAC clusters with Euclidean distance.
- Area covered by STAC clusters with Euclidean distance is more than the area of STAC clusters with Manhattan distance.
- Clusters in Manhattan distance become elongated to north south direction, whereas the shapes of clusters in Euclidean distance are more circular.

### 7.3.Discussion of the spatio-temporal crime prediction model

The core part of the crime prediction model is fitting data to the Box-Jenkins Arima model. All the values are updated and future values are gained according to the model. The limitation in this step is the size of data which only consisting one year. In most of the studies at least three year's data is used to detect the reflection of specific days and events. 5.23 number of crimes per day is found for future values.

The second part of the crime prediction model is simple spatial disaggregation approach. Approach starts with forming appropriate clusters to control the model and to predict the future values. As stated suitable clustering algorithm is selected as STAC with Manhattan and Euclidean distance application. As the crime prediction model is based on weekday values, data is divided according to daily intervals and all seven days are mapped and analyzed with STAC-Manhattan and STAC Euclidean algorithms. Forecasted values are evaluated in both of the algorithms to compare the accuracy of the clusters. Also, future values are assigned to clusters to give information about the high probable crime areas and number of crimes. According to analysis, STAC-Manhattan gives better results in terms of STRMSE of 1,08 whereas STAC-Euclidean gives 1.48. Results actually are not bad as predicting crime accurately is really a difficult subject. Also, STRMSE of every cluster per day are calculated and seen that bigger clusters having more number of incidents are deviating larger than the smaller clusters.

To sum up;

- The spatio-temporal crime prediction model in this study is adapted from Al-Madfai et. al. (2006). However, model is different from the original as explained in the previous sections. There are three important parameters which should be considered to adapt the model. Scale of the study area, the number of crime data and the time period of data. Scale of the study area and the number of crime data are considered in spatial analysis part. Also, time period is considered in temporal analysis of crime part. They applied the model to a city with nearly 17000 crime incidents. Hence, they generate the model according to these values. Our study area consists of two police precincts with nearly 1900 crime incidents. The model is adapted considering the difference between scale and number of data. Number of clusters generated in their study is higher than this study. The reason is because of the scale of the study area. Also, standard forecasting disaggregation weights are found by taking percentages of crime occurrences to the overall data per cluster. They used other methods but

their clusters include higher number of crimes. As most of the clusters has 0 observation per day in this study, results of the methods in their study will not be meaningful. Time period is three years in their study, whereas it is only one year data in this study. Hence, generated forecasting models are different.

- STRMSE of spatio-temporal crime prediction model are 1.08 and 1.48 for Manhattan and Euclidean distance applications respectively. Manhattan application gives better result than Euclidean application although STAC clusters with Manhattan distance has lower surface area. However, both error terms are not high for crime incidents.

- The number of clusters in weekdays indicates the distribution of crime incidents. If the number is high, incidents are more dispersed in that day. Police should consider this situation to prevent crime. Euclidean application of the model has 5 clusters both on Monday and Wednesday. At the end of week number of clusters which meaning concentrating in the area. In Manhattan application on Monday there are seven clusters generated. Also, at weekend number of clusters decrease. As a result, crime incidents in Monday are more dispersed in the study area. It gets difficult to control the area, when incidents are dispersed. Police should be aware of the situation to take precautions.

- The size of the clusters and the number of crime incident in the clusters are larger in Merkez Çankaya Police Precinct. The main reason is that more crime incidents are recorded at Merkez Çankaya Police precincts. However, according to the error terms for each cluster, at the same time predictions in Merkez Çankaya is less reliable than Bahçelievler police precincts.

- Clusters which cover small area indicates specific areas for crime. STAC clusters with Manhattan distance have smaller clusters. Places noticed by

small size clusters are Maltepe Bazaar, Olgunlar, Kocatepe Mosque, Konur Street, a market area in Bişkek Street, the intersection area of sixth and seventh streets, a school area in Anıttepe. These areas are mostly commercial and attractive areas for crime opportunities. The probability of occurrence of crime incidents in these areas are nearly 50%.

- On Saturday and Sunday number of predicted crime incidents increase in Merkez Çankaya. On Saturday people go to Merkez Çankaya region for shopping and number of pickpocket increases and on Sunday empty places are more prone to burglary and auto related theft.

- All clusters generated are compatible with crime theories under environmental criminology. Crime pattern theory explain the situation that most of the clusters are generated on main streets that people use for daily activities. Also routine activies theory explains the high number of crime incidents in Merkez Çankaya. Shopping is a routine activity that people in Ankara mostly prefer to go to Kızılay for shopping.

- Model is validated by seperating the last seven days as a control data. Euclidean distance metric clusters are found to be more accurate than Manhattan distance metric clusters.

### 7.4.Conclusion

In this thesis, a spatio-temporal crime prediction model is created. After the analysis, clusters per week day and the number of incidents in these clusters are predicted. Actually, these results represent sensitive areas to crime. Also, the number of incidents predicted indicates the level of sensitivity. Higher number of incidents predicted means the area is more prone to criminal activities. The solid results of this study is to determine these areas and the level of influence. To prevent crime before occurring is possible with identfying sensitive areas and reasons behind it. Reasons can be explained by crime theories under

environmental criminology. Generation of week day clusters is significant in terms of crime theories, as in each week day people have different daily activities. Routine activities theory states that people exposed to crime when they are doing daily activities. Each day has different cluster configuration as opportunities for some areas change according to the day of week. For example, many bazaar areas are open at only one day of week which provides opportunity for offenders in that area. If this bazaar area covered by a cluster on that day, the criminal activities can be reduced by taking precations in that area.

Police should utilize the model first by understanding the reason of clusters. Why the area covered by these clusters are attractive for offenders. This phase needs background information about the area. If area is known and identified in terms of land use, configuration of buildings, important organizations; it is possible to detect opportunities for crime in the crime triangle. Other phase is to take precations against crime. Situational crime prevention is helpful and effective if the structure of crime is detected.

Also, the methodology of this study can be used by police departments to be more informative about the future events. Crime is such an issue which is mobile and difficult to follow but makes patterns in the area concerning opportunities. Crime triangle theory states that there are three main necessities for crime; a motivated offender, absence of guardian and suitable target. Some areas are preparing the all the necessities and exposed to crime quiet often.

Developing such studies firstly needs eloborately collected data. Spatio-temporal crime prediction model can give better results if data on-hand consists more than one year. More data means better fitting of the model, more accurate clusters and hence more knowledge. When police understand the importance of crime analysis; hopefully, crime prevention in Turkey will be more effective. For example, spatio-temporal crime prediction model becomes useful in detecting real crime patterns, forecast the future values, take appropriate prevention measures and allocate resourses effectively. Thus, the rate of crime and number of offenders

decrease and in the other sense contribution in country's economy directly and indirectly.

In this study, a clustering algorithm STAC based on two difference metrics are implemented. Also, other clustering algorithms can be investigated with the same values in the same forecasting model and compared for future studies. Box-Jenkins Arima model is used to find forecasted values. Model can be developed by using other statistical forecasting models. Another future work can make a spatio-temporal crime prediction model with the same methodology to another study area and check the aplicability of the model.

## REFERENCES


Al-Madfai, H., Ivaha, C., Higgs, G.,Ware, A., Corcoran,J. (2007). “The Spatial Dissaggregation Approach to Spatio-Temporal Crime Forecasting,” International Journal of Innovative Computing, Information and Control, Vol. 3, Number 3.

Akpınar, E. “Using Geographic Information Systems in Analzing the Pattern of Crime Incidents and the Relationship Between Land Use and Incidents,”M.S thesis, Middle east Technical University, Turkey, 2005.

Ball, G.H., 'Data-analysis in the social sciences: What about the details?', Proceedings of the Fall Joint Computer Conference, 27, 533-559 (1966).

Boba, R. (2005) Crime Analysis and Crime Mapping, Sage, USA.

Bowerman, B. L. & O’Connell, R. T. (1993). Forecasting and Time Series: an Applied Approach, Pacific Grove, CA: Duxbury

Brantingham, P.J and Brantingham, P.L. (Eds.) (1981). Environmental Criminology, Beverly Hills: Sage.

Chainey, S. And Ratcliffe, J. (2005) GIS and Crime Mapping, John Wiley, England.

Cohen, J., Gorr, W.L., and Olligschlaeger, A. (2007), "Leading Indicators and Spatial Interactions: A Crime Forecasting Model for Proactive Police Deployment" Geographical Analysis, Vol.39 (1),pp. 105–127.

Corcoran J., Wilson I.D., and Ware J.A. (2003), “Predicting the Geo-Temporal Variations of Crime and Disorder”, International Journal of Forecasting, Vol 19, 2003, pp 623-634.

Deadman, D. (2003). “Forecasting Residential Burglary,” International Journal of Forecasting, Special Section on Crime Forecasting, Vol. 19, pp. 567-578.

Everitt, B. (1974) Cluster Analysis, Heinemann, London.

Felson, M., & Poulsen, E. (2003). Simple indicators of crime by time of day. International Journal of Forecasting, Vol.19, pp.595-602.

Felson,M. and R.V. Clarke(1998). Opportunity Makes the Thief. Crime Detection and Prevention Series,Paper 98. Police research Group. London: Home Office.

Gorr, W.L. and Harries, R. (2003). “Introduction to Crime Forecasting,” International Journal of Forecasting, Special Section on Crime Forecasting, Vol. 19, pp. 551-555.

Gorr W.L., Olligschlaeger A.(1997).”Spatio-Temporal Forecasting of Crime: Application of Classical and Neural Network Methods.” Working Papers of The Heinz School, Carnegie Mellon University.

Groff, E.R. and LaVigne, N.G.(2002). Forecasting the future of predictive crime mapping, NY

Grubesic, T.H. (2006). “On the Application of Fuzzy Clustering for Crime Hot Spot Detection,” Journal of Quantitative Criminology, Vol. 22, No.1.

Harries, R. (2003). Modelling and predicting recorded property crime trends in England and Wales--a retrospective.International Journal of Forecasting, 2003, vol. 19, issue 4, pages 557-566.

Hirschfield, A. and Bowers, K. (2001) Mapping and Analysing Crime Data: Lessons from Research and Practice, Taylor and Francis, New York.

Kaufman, L., and Rousseeeuw, P. (1990). Finding Groups in Data:An Introduction to Cluster Analysis, John Wiley,New York.

Levine, N. (2002), Crimestat: A Spatial Statistics Program for the Analysis of Crime Incident Locations, Ned Levine and Associates and the National Institute of Justice, Washington, DC.

National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, Thomas F. Rich,"The Use of Computerized Mapping in Crime Control and Prevention Programs.",July 1995.

Openshaw, S. (1998) Building automated geographical analysis and explanation machines. In Longley et al. (eds.) Geocomputation: A Primer. Chichester: Wiley: 95-115

Olligschlaeger, A. (1997). Artificial neural networks and crime mapping, NY

Pena, D. Tiao, G.C. Tsay, R.S. (2001) A Course in Time Series Analysis, John Wiley, Canada

Phillips, P.C.B. and P.Perron (1988): Testing for a Unit Root in Time Series Regression. Biometrica 75, 335-346

Ratcliffe J.H. (2004) "the hot spot matrix :a framework for the spatio temporal targeting of crime reduction police practice and research," Vol.5(1) pp.5-23.

Vann, B.I. and Garson, D. (2003) Crime Mapping, Peter Lang Publishing, New York.

Weisburd, D., Maher, L and Sherman L. W. (1992). Contrasting Crime General and Crime Specific Theory: The Case of Hot-Spots of Crime. Advances in Criminological Theory. Vol.4, 45-70.NJ:Transaction Press.

Web1: http://tr.wikipedi.org/wiki/Ankara (visited on 03.02.2007)

Web2: http://www.cankaya_bld.gov.tr/cankaya.asp, "Çankaya Hakkında" (visited on 03.02.2007)

Web3: http://www.ojp.usdoj.gov/nij/178919, National Institute of Justice, Office of Justice Programs, U.S. Department of Justice,"Mapping crime: principle and practice,"1999. (visited on 05.05.2007)

Web4:http://aic.gov.au/publications/crm/crm030.pdf, Australian Institute of Criminology. "Is crime predictable?" (visited on 03.03.2007)

Web5:http://www.britannica.com/eb/topic-178294/econometric-model, "Econometric model" (visited on 04.03.2007)

Web6:http://www.itl.nist.gov/div898/handbook/, "NIST/SEMATECH eHandbook of Statistical Methods"(visited on 13.05.2007)

Web7: http://www.autobox.com, "Autobox user's guide"(visited on 15.05.2007)

Web8: http://www.ncss.com/download.html#Manuals in Box, G. E. P., Jenkins G. M. (1976). Time Series Analysis Forecasting and Control, San Francisco. (visited on 09.06.2007)

## INTERVIEWS

Interview with Rana Sampson (2006) international crime consultant, the former director of public safety for the University of San Diego.

Interview with a police officer (2007) in Bahçelievler Police Station.